Department of Medicine

Integrating -*Omics* For Studying Functional Role Of Ulcerative Colitis Risk Associated Loci.

Lauma Ramona

"This dissertation is submitted for the degree of Doctor of Philosophy"

Project Supervisors: Dr. Matthew Robinson and Professor Arthur Kaser

July 2019

Declaration

I declare that all work present have been composed entirely by the author and not been previously published or submitted for degree. All sources of information other than author have been fully referenced and experimental help by those others than the author have been appropriately acknowledged. Work does not exceed the prescribed word limit.

Lauma Ramona

July 2019

Abstract

Background: Ulcerative Colitis is chronic inflammatory condition of unknown etiology. Genome Wide Association Studies have successfully identified large number of UC risk associated loci, majority of which are located in non-protein-coding DNA regions and been showed to be enriched within regulatory elements, such as enhancers. However, the function of these UC risk associated variants is still unknow.

Aim: To delineate the functional role of GWAS risk associated loci in UC relevant cell types.

Method: We assessed chromatin activity (ATAC seq) and transcriptional behavior (RNA seq) of primary cell types extracted from intestinal biopsies and blood from diseased and healthy participant. Next, to pinpoint the mechanistic of how UC associated loci contributes to disease risk, we intersected our disease and cell type specific differential expression and differential chromatin accessibility data with GWAS dataset.

Results: Unfortunately, due to technical and financial reasons we failed to reach the target sequencing depth for both ATAC seq and RNA seq experiments. In addition, when combined with very low participant numbers, our data sets were not strong enough to reliably identify the functional role of GWAS variants. However, for practice, we proceeded with slightly simplistic proximity-based modeling and showed that intersecting the 3 -*omics* studies allowed us to identify 10 regions where the lowest *p*-value associated SNP was in proximity to differentially expressed gene and differentially accessible chromatin region.

Conclusion: We were able to compare the first time the expression levels and chromatin conformation in purified immune cell populations from intestinal tissue and peripheral blood. Unfortunately, due to poor experimental design this study was markedly underpowered and any findings from RNA seq and ATAC seq experiments should be further validated before any biological conclusions are made or used for reliable prediction of functional role of UC associated risk variants.

Acknowledgement

The author wishes to acknowledge MedImmune and the Cambridge NIHR BRC for their financial support.

The author wishes to thank the Project Supervisors for their guidance and encouragement, and Josquin Nys for his patient and support in laboratory work.

My sister. Without her I would be lost.

Contents

LIST OF TABLES 9

LIST OF FIGURES 11

ABBREVIATIONS 14

1. INTRODUCTION 18

1.1 Ulcerative Colitis 19
1.2 Clinical Presentations And Available Treatments 20
1.3 Pathogenesis Of UC 22
1.4 Genetics Studies Of Common Disease 24
1.5 Difficulties With GWAS Interpretation 25
1.6 GWAS Risk Variants Enrichment On *Cis*-Regulatory Regions 26
1.7 Determining Activity Of Regulatory Regions 28
1.8 Computational Models For Functional Interpretation Of GWAS Data 30
1.8.1 *GWAS SNP Enrichment Analysis* *31*
1.8.2 *Colocalization Analysis* *32*
1.9 Disease Specific Cell Type(s) And Their Rolle In GWAS Interpretation 34
1.9.1 *Intestinal Epithelium* *35*
1.9.2 *Intraepithelial Lymphocytes* *35*
1.9.3 *Lamina Propria Residing T Cells* *36*
1.9.4 *Lamina Propria Residing B Cells* *37*

2. AIMS 39

3. MATERIALS AND METHODS 41

3.1 Recruitment Of Healthy Donors And UC Patients 42
3.2 Volunteer Demographics 43
3.3 Purification Of Individual Cell Types 46
3.3.1 *Peripheral Blood Mononuclear Cell Isolation* *46*
3.3.2 *Lamina Propria And Intraepithelial Cell Separation* *47*
3.3.3 *Cell Staining* *47*
3.4 Flow Cytometry 48
3.4.1 *Cell Phenotyping* *48*
3.4.2 *Cell Sorting* *48*
3.5 RNA Sequencing Library Construction And Quality Control 51
3.5.1 *RNA Extraction And Quality Control* *51*
3.5.2 *Smarter Stranded Total RNA Seq Library Construction, Quantification And Quality Control* *51*
3.5.3 *Library Multiplexing, Pooling and Quality Control* *52*
3.6 ATAC Sequencing Library Construction And Quality Control 52
3.6.1 *Nuclear Isolation And Chromatin Tagmentation* *54*
3.6.2 *ATAC Seq Library Construction And Quantification* *54*
3.7 Data Analysis 56
3.8 List Of Published Software Packages Used In This Study 57

4. COMPARISON AND VALIDATION OF COMMERCIALLY AVAILABLE ANTI-GPR15 ANTIBODIES FOR USE IN FLOW CYTOMETRY 59

4.1 Introduction 60
4.2 Aim 61

4.3 MATERIALS AND METHODS 62
4.3.1 *Generation Of Jump-In T-REx HEK293 Cell Lines Which Over-Expresses Human GPR15* 62
4.3.1.1 *Stable Transfections* 63
4.3.1.2 *Transient Transfections* 63
4.3.1.3 *Plating For Antibody Staining* 63
4.3.2 *Peripheral Blood Mononuclear Cell Isolation From Blood Cones* 64
4.3.3 *Antibody Staining* 64
4.3.4 *Assessment Of GPR15 Expression At mRNA Level* 66
4.3.5 *Flow Jo Analysis* 66
4.4 RESULTS 67
4.4.1 *Comparison Of Commercially Available Anti-GPR15 Antibody Performance Using Genetically Engineered Jump-In HEK293 Cell Lines.* 67
4.4.2 *Assessment Of MAB3654 Staining On Peripheral Blood T Cells* 70
4.4.3 *MAB3654 Staining Optimization For Flow Cytometry Using Blood Resident T Cells* 72
4.4.4 *MAB3654 Staining Validation Using Blood Resident T Cells GPR15* 77
4.5 DISCUSSION 79

5. ASSESSMENT OF DIFFERENCES IN TRANSCRIPTIONAL PROFILE BETWEEN THE HEALTHY CONTROLS AND UC PATIENTS 82

5.1 INTRODUCTION 83
5.2 AIM 84
5.3 MATERIALS AND METHODS 85
5.3.1 *Sequencing Design* 85
5.3.2 *RNA Seq Data Analysis Pipeline* 85
5.3.2.1 *Pre-Processing Of Raw Sequencing Data* 85
5.3.2.2 *Downstream Analysis* 86
5.3.2.2.1 Creation Of Gene-Level Count Datasets 86
5.3.2.2.2 Gene And Sample Sub-Setting 86
5.3.2.2.3 Sample Quality 87
5.3.2.2.4 Call For Differential Expression 87
5.3.2.2.5 Pathway Analysis 88
5.3.2.2.6 Overlay With Published Literature 88
5.4 RESULTS 90
5.4.1 *Determining Disease Specific Change In Expression Profiles In Purified CD19⁺ B Cells And CD4⁺ T_{EM} Immune Cell Populations From Peripheral Blood And Sigmoid Colon* 90
5.4.2 *In-Silico Validation Of Genes Identified As Differentially Expressed Between The UC Patients And Healthy Individuals* 97
5.4.3 *Determining Biological Meaning Behind The Disease Specific Change In Expression Profiles* 99
5.4.4 *Determining Anatomical Location Specific Change In Expression Profiles In Purified CD19⁺ B Cells And CD4⁺ T_{EM} Immune Cell Populations.* 104
5.5 DISCUSSION 108

6. COMPARISON OF CHROMATIN ACCESSIBILITY BETWEEN THE HEALTHY VOLUNTEERS AND UC PATIENTS 112

6.1 INTRODUCTION 113
6.2 AIM 114
6.3 MATERIALS AND METHODS 115
6.3.1 *ATAC Seq Library Sequencing* 115
6.3.2 *ATAC Seq Data Analysis Pipeline* 115
6.3.2.1 *Pre-Processing Of Raw Sequencing Data* 115

6.3.2.2 *Downstream Analysis* *116*
6.3.2.2.1 Peak Consolidation *116*
6.3.2.2.2 Peak Annotation *117*
6.3.2.2.3 Counts Matrix Generation Based On Peak Coordinates In Consolidated Peaksets..... 118
6.3.2.2.4 Sample Selection And Data Quality Control *119*
6.3.2.2.5 Call For Differential Accessibility *119*
6.4 RESULTS 120
6.4.1 *Determining Chromatin Landscape In Purified Cell Populations From Peripheral Blood And Sigmoid Colon Lamina Propria And Intraepithelial Layers From Healthy Volunteers And UC Patients...* *120*
6.4.2 *Determining Differences In Chromatin Accessibility In Purified Cell Populations From Peripheral Blood And Sigmoid Colon Lamina Propria And Intraepithelial Layers From Healthy Volunteers And UC Patients*..... *124*
6.5 DISCUSSION 126

7. COMBINATORIAL ANALYSIS OF FUNCTIONAL GENOMICS AND GENETICS DATA 129

7.1 INTRODUCTION 130
7.2 AIM 131
7.3 MATERIALS AND METHODS 132
7.3.1 *RNA Seq Data Correlation With ATAC Seq* Data *132*
7.3.2 *Calculation Of Genes And Chromatin Regions Enrichment Within GWAS Risk Loci* *133*
7.3.3 *Integration Of UC Risk Associated SNPs With RNA Seq And ATAC Seq Data* *135*
7.4 RESULTS 137
7.4.1 *Estimating How Much Of The Chromatin Regulatory Landscape Can Be Inferred From Gene Expression And Vice Versa* *137*
7.4.2 *Enrichment Of Disease Associated Genes And Chromatin Regions Within The GWAS Risk Loci Associated With Immune Mediated Diseases Or Traits* *143*
7.4.2.1 *Testing For Disease State Specific Expression Enrichment* *143*
7.4.2.2 *Testing For Disease State Specific Chromatin Conformation Enrichment* *146*
7.4.2.3 *Testing For Cell Linage Specific Expression Enrichment* *146*
7.4.3 *Integrating GWAS, ATAC Seq And RNA Seq Data To Predict Molecular Mechanisms By Which UC Risk Associated Variants Might Contributes To The Studied Phenotype.* *148*
7.5 DISCUSSION 151
7.5.1 *Relationship Between Expression And Accessibility* *151*
7.5.2 *Disease And Trait Associated Locus Enrichment For DEG And DA Regions* *152*
7.5.3 *Integration Of Functional Genomics Data With GWAS Identified UC Associated Risk Variants* *155*

8. DISCUSSION 161

9. REFERENCES 166

Appendix 1 – RNA SEQUENCING OPTIMIZATION 199

Appendix 2 – *IN SILICO* PREDICTIONS OF INDIVIDUAL RNA SEQ LIBRARY SEQUENCING PERFORMANCE 204

Appendix 3 – EXTENDED RNA SEQ DATA QC 206

Appendix 4 – CHALLENGES ENCOUNTERED DURING RNA SEQ ANALYSIS 220

Appendix 5 – POWER CALCULATION FOR RNA SEQ EXPERIMENTS 224

Appendix 6 – EXTENDED ATAC SEQ DATA QC 227

Appendix 7 – COUNT NORMALIZATION METHODS AND THEIR APPLICATION IN ATAC SEQ ANALYSIS 236

Appendix 8 – CHALLENGES ENCOUNTERED DURING ATAS SEQ ANALYSIS 239

List Of Tables

Table 3.1 PARTICIPANT DEMOGRAPHICS AND SAMPLES USED IN THE FINAL ANALYSIS. 44

Table 3.2 AGE, SEX AND SAMPLE NUMBERS AND RATIOS FOR EACH INDIVIDUAL POPULATION USED FOR DIFFERENTIAL ACCESSIBILITY OR DIFFERENTIAL EXPRESSION ANALYSIS. 45

Table 3.3 LIST OF ALL ANTIBODIES USED DURING THE SEQUENCING STUDY. 48

Table 3.4 LIST OF ALL REAGENTS REQUIRED FOR ATAC SEQ LIBRARY CONSTRUCTION. 53

Table 3.5 ATAC SEQ BUFFER RECIPES. 53

Table 3.6 LIST OF ATAC SEQ BARCODE NUCLEOTIDE SEQUENCES (Buenrostro *et al.*, 2013)**.** 54

Table 4.1 LIST OF ALL REAGENTS REQUIRED FOR JUMP-IN T-REX HEK293 TRANSFECTION EXPERIMENTS. 62

Table 4.2 SUMMARY OF BUFFER RECIPES NEEDED FOR JUMP-IN T-REX HEK293 TRANSFECTION EXPERIMENTS. 62

Table 4.3 LIST OF ALL ANTIBODIES USED IN ANTI-GPR15 ANTIBODY EVALUATION EXPERIMENTS. 66

Table 5.1 NUMBER OF DEG IDENTIFIED IN SELECTED IMMUNE CELL POPULATIONS FROM HEALTHY SUBJECTS AND UC PATIENTS WITH DIFFERENT EXTENT OF DISEASE (*for* n_{donor} *for each subset please see Table 3.2* 91

Table 5.2 LIST OF DEG IDENTIFIED IN SELECTED IMMUNE CELL POPULATIONS FROM HEALTHY SUBJECTS AND UC PATIENTS WITH DIFFERENT EXTENT OF DISEASE. 92

Table 5.3 IPA PATHWAY ENRICHMENT ANALYSIS OF GENES DIFFERENTIALLY EXPRESSED IN BLOOD $CD19^+$ B CELL (CONTROL VS UCI), LPL $CD19^+$ B CELL (CONTROL VS UCN) AND LPL $CD4^+$ T_{EM} (CONTROL VS UCI) COHORTS. 100

Table 5.4 PUBMED LITERATURE SEARCH FOR DEG IDENTIFIED IN STUDY. 103

Table 5.5 LIST OF TOP 10 DEG (BY LOG2FOLDCHANGE) IDENTIFIED IN SELECTED IMMUNE CELL POPULATIONS FROM PERIPHERAL BLOOD AND SIGMOID COLON LAMINA PROPRIA. 105

Table 5.6 IPA PATHWAY ENRICHMENT ANALYSIS OF GENES DIFFERENTIALLY EXPRESSED IN CD19 B CELL (BLOOD *VS* LPL) AND $CD4^+$ T_{EM} (BLOOD *VS* LPL) COHORTS. 106

Table 6.1. LIST OF DIFFERENTIALLY ACCESSIBLE REGIONS 125

Table 7.1 SPEARMAN CORRELATION BETWEEN *A.* DEG AND DA, *B.* DEG AND GLOBAL CHANGES IN THE CHROMATIN PROFILE, *C.* DA AND GLOBAL CHANGES IN THE EXPRESSION PROFILE AND *D.* GLOBAL CHANGES IN THE CHROMATIN PROFILE AND GLOBAL CHANGES IN THE EXPRESSION PROFILE. 139

Table 7.2 IMMUNE-DISEASE OR TRAIT ASSOCIATED VARIANT ENRICHMENT FOR PEAKS AND GENES IDENTIFIED AS SIGNIFICANTLY DIFFERENT IN CELL POPULATIONS VARYING BY DISEASE STATE OR ANATOMICAL LOCATION. 145

Table 7.3 LIST OF GENES THAT WERE IDENTIFIED AS DIFFERENTIALLY EXPRESSED IN B CELLS FROM PERIPHERAL BLOOD COMPARED TO THEIR GUT COUNTERPARTS AND ENRICHED IN BOTH UC AND CD RISK ASSOCIATED LOCUS. 147

Table 7.4 LIST OF DEG AND DA WHICH EITHER FELL INTO OR WERE IN PROXIMITY TO UC ASSOCIATED RISK LOCUS. 149

List Of Figures

Figure 1.1 GEOPOLITICAL REPRESENTATION OF WORLDWIDE INCIDENCE OF UC EXPRESSED PER 100'000 PERSON-YEARS (Ng *et al*. 2020). **20**

Figure 1.2 ENDOSCOPIC OUTLOOK OF B. MILDLY, C. MODERATELY AND D. SEVERELY AFFECTED INTESTINAL TISSUE OF UC PATIENTS IN COMPARISON TO A. HEALTHY SUBJECT (Kobayashi *et al.,* 2020). **21**

Figure 1.3 RISK FACTORS CURRENTLY BELIEVED TO HAVE SOME POSSIBLE IMPLICATIONS IN DEVELOPMENT OF IBD (Turpin *et al.,* 2018). **23**

Figure 1.4 *A.* GRAPHICAL REPRESENTATION OF ATAC SEQ WORKING PRINCIPLES, *B.* TIME AND INPUT MATERIAL REQUIREMENTS DEPENDING ON SELECTED METHOD FOR OPEN-CHROMATIN ANALYSIS *(Buenrostro et al., 2013).* **29**

Figure 3.1 SUMMARY OF CELL POPULATIONS AND THEIR SORTING MARKERS. **46**

Figure 3.2 AN EXAMPLE OF FACS GATING STRATEGY USED TO PURIFY INDIVIDUAL CELL POPULATION FROM *A.* BLOOD, *B.* LPL AND *C.* IEL SAMPLES, ALL OBTAINED FROM A SINGLE DONOR. **51**

Figure 4.1 DIAGRAMS SHOWING THE STEP BY STEP STAINING STRATEGY FOR EACH INDIVIDUAL ANTI-GPR15 ANTIBODY COMPARISON, VALIDATION AND EVALUATION EXPERIMENT. **65**

Figure 4.2 COMPARISON OF COMMERCIALLY AVAILABLE ANTI-GPR15 ANTIBODIES *($n_{experiment} = 2$, $n_{transfection\ attempts} = 2$)*. **69**

Figure 4.3 FLOW CYTOMETRY PLOTS SHOWING THE ANTI-GRP15 ANTIBODY MAB3654 STAINING PERFORMANCE ($n_{experiment} = 2$, $n_{donor} = 4$). **71**

Figure 4.4 FLOW CYTOMETRY PLOTS SHOWING THE *A.* INITIAL GATING STRATEGY, *B.* MAB3654 PRIMARY AND *C.* GOAT-ANTI-MICE SECONDARY ANTIBODY TITRATION, *D.* TAG COMPARISON AND *E.* MAB3654 INCUBATION TIME COMPARISON. **76**

Figure 4.5 MAB3654 STAINING VALIDATION WITH *A.* qPCR ($n_{experiment} = 1$, $n_{donor} = 4$) **AND *B.* IgG 2b ISOTYPE CONTROL** ($n_{experiment} = 1$, $n_{donor} = 1$)**.** **78**

Figure 6.1. GRAPHICAL ILLUSTRATION OF STEPS INVOLVED IN PEAK FILTERING AND UNIFIED COORDINATE ESTABLISHMENT. **117**

Figure 6.2 VENN DIAGRAMS SHOWING OPEN CHROMATIN REGIONS THAT WERE UNIQUE TO OR SHARED BETWEEN DIFFERENT DISEASE STATES IN A CELL TYPE AND ANATOMICAL LOCATION SPECIFIC MANNER ($n_{phenotype} = 30$; *For* n_{donor} *please see Table 3.2).* **121**

Figure 6.3 SCATTER PLOTS ILLUSTRATING THE RELATIONSHIP BETWEEN THE PEAK COUNT, SAMPLE NUMBER AND GEOMETRIC MEAN OF COUNTS ASSOCIATED WITH THE SAME PHENOTYPE ($n_{phenotype} = 30$)**.** **122**

Figure 6.4 GENOMIC ALIGNMENT OF ACCESSIBLE REGIONS. **123**

Figure 7.1 OUTLINE OF THE ALGORITHM USED FOR CALCULATING ENRICHMENT IN REGIONS SURROUNDING DISEASE OR TRAIT ASSOCIATED FOCAL SNPS (Raine *et al.*, 2015). 135

Figure 7.2 VISUAL REPRESENTATION OF RELATIONSHIP BETWEEN GENE EXPRESSION AND PROMOTER ACCESSIBILITY IN *A.* BLOOD CD19$^+$ B CELL, *B.* LPL CD19$^+$ B CELL AND *C.* LPL CD4 T_{EM} CELLS VARYING BY DISEASE STATE AND *D.* CD4$^+$ T_{EM} VS CD19$^+$ B CELLS 141

Figure A1.1 RAW RNA SEQUENCING READ ALIGNMENT TO THE HUMAN GENOME EXPRESSED AS PERCENTAGE OF TOTAL READS ($n_{sample} = 13$). 200

Figure A1.2 PERCENTAGE OF UNMAPPED READS THAT EITHER DID OR DID NOT REALIGN AGAINST ANY OF OTHER GENOMES TESTED (n $_{sample}$ = 13). 200

Figure A1.3 RELATIONSHIP BETWEEN THE RNA QUALITY AND PERCENTAGE OF READ REMAPPING TO THE VIRAL, BACTERIAL AND HUMAN GENOME (n $_{sample}$ = 13). 201

Figure A1.4 GENOMIC ALIGNMENT OF MAPPED READS. 202

Figure A2.1 RELATIONSHIP BETWEEN THE RNA SEQUENCING LIBRARY ALIGNMENT AND EARLY SAMPLE/LIBRARY METRICS (n $_{sample}$ = 13). 204

Figure A3.1 RNA LIBRARY SEQUENCING DEPTH AND ALIGNMENT *($n_{sample} = 92$).* 206

Figure A3.2 RELATIONSHIP BETWEEN THE RNA SEQUENCING LIBRARY ALIGNMENT AND EARLY SAMPLE/LIBRARY METRICS *($n_{sample} = 92$).* 207

Figure A3.3 GENOMIC ORIGIN OF ALIGNED READS *($n_{sample} = 92$).* 207

Figure A3.4 EXPLORATORY ANALYSIS ($n_{samples} = 94$). 209

Figure A3.5 PRINCIPAL COMPONENT ANALYSIS ON A. BLOOD CD4$^+$ T_{EM}, B. BLOOD CD19$^+$ B, C. LPL CD4$^+$ T_{EM} AND D. LPL CD19$^+$ B CELL POPULATIONS. 215

Figure A3.6 GRAPHICAL REPRESENTATION OF PC1/PC2 FOR CD4$^+$ T_{EM} AND CD19$^+$ B CELLS FROM PERIPHERAL BLOOD AND SC LPL. 215

Figure A3.7. EVALUATION OF CELL POPULATION PURITY BASED ON CELL SURFACE MARKER EXPRESSION. 216

Figure A3.8 REPRESENTATIVE COUNTS DISTRIBUTION FOR A. LOW EXPRESSED GENE AND B. RELATIVELY HIGHER EXPRESSED GENE. 218

Figure A3.9. P-VALUE VISUALIZATION BEFORE AND AFTER CORRECTION FOR THE STANDARD DEVIATION FOR A. CD4$^+$ T_{EM} (Blood *vs* LPL) AND B. CD19$^+$ B CELLS (Blood *vs* LPL) COMPARISONS. 219

Figure A5.1 POWER CURVE FOR CD4$^+$ T CELL DATA. 225

Figure A5.2 POWER CURVE FOR CD19$^+$ B CELL DATA. 226

Figure A6.1 ATAC LIBRARY SEQUENCING DEPTH (n $_{sample}$ = 183)**.** 227

Figure A6.2 ATAC SEQUENICNG LIBRARY ALIGNMENT TO HUMAN GENOME EXPRESSED AS PERCENTAGE OF TOTAL RAW READS (n $_{sample}$ = 183). 228

Figure A6.3 PERCENTAGE OF TOTAL PROCESSED READS THAT MAPPED ON MITOCHONDRIAL GENOME (n $_{sample}$ = 183)**.** 229

Figure A6.4 READ COUNT PER SAMPLE POST FILTERING THAT WAS USED FOR PEAK CALLING (n $_{sample}$ = 183)**.** 229

Figure A6.5 PERCENTAGE OF FRACTION OF READS IN PEAKS (n $_{sample}$ = 183). 230

Figure A6.6 HEATMAP REPRESENTING SAMPLE-TO-SAMPLE RELATIONSHIPS *(n $_{samples}$ = 158)*. 232

Figure A6.7 P-VALUE HISTOGRAMS ILLUSTRATING CONSERVATIVE (HILL-SHAPED), UNIFORM AND ANTI-CONSERVATIVE DISTRIBUTIONS. 233

Figure A7.1 SPERAMAN CORRELATION LOOKING AT THE RELATIONSHIP BETWEEN THE TOTAL LIBRARY SIZE AND READS IN PEAKS (*n $_{samples}$ = 158*). 236

Figure A7.2 VENN DIAGRAM REPRESENTING THE NUMBER OF DIFFERENTIALLY ACCESSIBLE REGIONS IDENTIFIED FROM THE SAME DATA SET NORMALIZED BY EITHER MEDIAN RATIO METHOD OR SCALED TO FULL LIBRARY SIZE. 237

Figure A7.3 KEGG PATHWAY ANALYSIS OF CHROMATIN REGIONS IDENTIFIED AS DA BY MEDIAN RATIO METHOD. 238

Abbreviations

APCs - Antigen Presenting Cells

ATAC seq - Assay For Transposonase Accessible Chromatin Using Sequencing

ATG16L1 - Autophagy Related 16 Like 1

BMDCs - Bone Marrow-Derived Dendritic Cells

BMI - Body Mass Index

BSA - Bovine Serum Albumin

CASP8 - Caspase 8

CASP8AP2 - Caspase 8 Associated Protein 2

CBX3 - Chromobox 3

CCR7 - CC-Chemokine Receptor 7

CD - Crohn's Disease

CD4 - Cluster Of Differentiation 4

CD45 - Cluster Of Differentiation 45

CD62L - Cluster Of Differentiation 62 L

CD69 - Cluster Of Differentiation 69

CD71 - Cluster Of Differentiation 71

Chip-seq - Chromatin Immunoprecipitation Combined With Sequencing

DA - Differential Accessibility

DAR - Differentially Accessible Regions

DEG - Differential Gene Expression

DENND1B - DENN Domain Containing 1B

DHS - DNase I Hypersensitivity Sites

DMEM - Dulbecco's Modified Eagle's Medium

DNase I - Deoxyribonuclease I

DO - Disease Ontology

DTT - DL-Dithiotheritol

DUOX2 - Dual Oxidase 2

EDTA - Ethylenediaminetetraacetic Acid

ENCODE - Encyclopaedia Of DNA Elements

eQTL - Expression Quantitative Trait Loci

FACS - Fluorescence Activated Cell Sorting

FAM53B - Family With Sequence Similarity 53 Member B

FcgR1 - High Affinity Immunoglobulin Gamma Fc Receptor I

FRiP - Fraction Of Reads In Peaks

GATA3 - GATA Binding Protein 3

gDNA - Genomic DNA

GFP - Green Fluorescence Protein

GO - Gene Ontology

GTEx - Genotype-Tissue Expression Project

GWAS - Genome-Wide Association Studies

HBSS - Hank's Balanced Salt Solution

HELZ2 - Helicase With Zinc Finger 2

IBD - Inflammatory Bowel Disease

IDH2 - Isocitrate Dehydrogenase

IEL - Intraepithelial Lymphocytes

IgA - Immunoglobulin A

IgM - Immunoglobulin M

IIBDGC - International IBD Genetics Consortium

IL-13 - Interleukin 13

IL-15 - Interleukin 15

IL-17A - Interleukin 17A

IL23R - Interleukin 23 Receptor

IL-4 - Interleukin 4

IL-5 - Interleukin 5

IL-6 - Interleukin 6

IPA - Ingenuity Pathway Analysis

JUP - Junction Plakoglobin

KSR2 - Kinase Suppressor of Ras 2

LD - Linkage Disequilibrium

LPL - Lamina Propria

LPS - Lipopolysacharide

MCV - Mean Corpus Volume

MF - Macrophages

MHC - Major Histocompatibility Complex

MHC II - Major Histocompatibility Complex Class II

MIER3 - MIER Family Member 3

MS - Multiple Sclerosis

MUC2 - Mucin 2

NAB1 - NGFI-A Binding Protein 1

NHGRI – EBI - National Human Genome Research Institute - European Bioinformatic Institute

NOD2 - Nucleotide-Binding Oligomerization Domain Containing 2

NODs - Nucleotide Binding Oligomerization Domain-Like Receptors

NOL4L - Nucleolar Protein 4 Like

PBMCs - Peripheral Blood Mononuclear Cell

PC - Principal Component

PCA - Principal Component Analysis

pIgR - Polymeric Ig Receptor

PRRs - Pattern-Recognition Receptors

RhArt - Rheumatoid Arthritis

RIN - RNA Integrity Number

RPKM - Reads Per Kilobase Million

rRNA - Ribosomal RNA

S1PR1 - Sphingosine 1-Phosphate Receptor Type 1

SC - Sigmoid Colon

SEMA4A - Semaphorin 4A

SNP - Single Nucleotide Polymorphism

SNX27 - Sorting Nexin Family Member 27

SPRED2 - Sprouty Related EVH1 Domain Containing 2

T1D - Type 1 Diabetes

T2D - Type 2 Diabetes

T_{CM} - T Central Memory

T_E - T Effector Cells

T_{EM} - T Effector Memory Cells

T_{EMRA} - T Terminal Effector Memory

TF - Transcription Factor

TFBSs - Transcriptional Factor Binding Sites

T_H - T Helper Cells

TI - terminal ileum

TLRs - Toll Like Receptors

T_M - T Memory Cells

TNF-α -Tumour Necrosis Factor Alpha

TRAFD1 - TRAF-Type Zinc Finger Domain Containing 1

TSS - Transcription Start Site

T_{TRM} - Tissue Resident Memory T Cells

TXNIP - Thioredoxin Interacting Protein

UC - Ulcerative Colitis

UC_i - UC With Inflammation In The Sigmoid Colon

UC_n - UC Without Inflammation In The Sigmoid Colon

XBP1 - X-Box Binding Protein 1

1. Introduction

1.1 Ulcerative Colitis

Ulcerative colitis (UC) is chronic, idiopathic inflammatory disorder characterized by relapsing inflammation within the gastrointestinal tract. UC is one of two most common forms of Inflammatory Bowel Disease (IBD). The other main form of IBD is a disorder known as Crohn's disease (CD) (Mozdiak *et al.,* 2015).

In 2012 *Molodecky et al.* reported the highest IBD incidence rates among 20 - 29 year old's (Molodecky *et al.*, 2012). However, there are growing number of publications highlighting an increasing incidence of UC in pediatric (Sýkora *et al.,* 2018) and elderly (Zammarchi *et al.,* 2020) populations. It was reflected in recent report by National Institute for Health and Care Excellence showing the highest UC incidence rates among 15 - 25 year old's, with second smaller peak between 55 - 65 year old's (NICE guidelines 2019).

With better therapies becoming available, UC have transitioned from disease being associated with high mortality rates into condition characterized by lifelong disability (GBD 2017 Inflammatory Bowel Disease Collaborators 2020). Being a lifelong condition, UC can lead to emotional distress, anxiety and depression, particularly among young adults. Moreover, with the standard care being extremely costly, UC is becoming increasingly larger economic burden on the healthcare system. Associated healthcare costs of IBD in the UK in 2013 was £470 million (The IBD Standards Group, 2013) and *Burisch et al* reported that direct healthcare costs of IBD in Europe was approximately €4.6 - €5.6 billion per year (Burisch *et al.*, 2013).

The prevalence of UC is highest in Northern America, Northern Europe and UK (also referred as regions with high socio-demographic index; Figure 1.1) (Ananthakrishnan, 2015; Ng *et al.*, 2017). The estimated prevalence of UC in the UK is 1 in 420 (Crohn's and Colitis UK, no date). Although UC incidence rates are stabilizing in countries associated with historically high UC occurrence, new cases have been rising in former low incidence areas, making UC emerge as a global disease (Loftus, 2004; Kaplan and Ng, 2017). The exact reason behind this phenomenon is still unknown, however it has been hypothesized that introduction western-like-diet (Lerner and Matthias 2015)

and increase in hygiene might be some of the risk factors contributing to observed increase. Nevertheless, the availability of better diagnostic tools and standardized diagnosis in former low incidence regions should not be excluded.

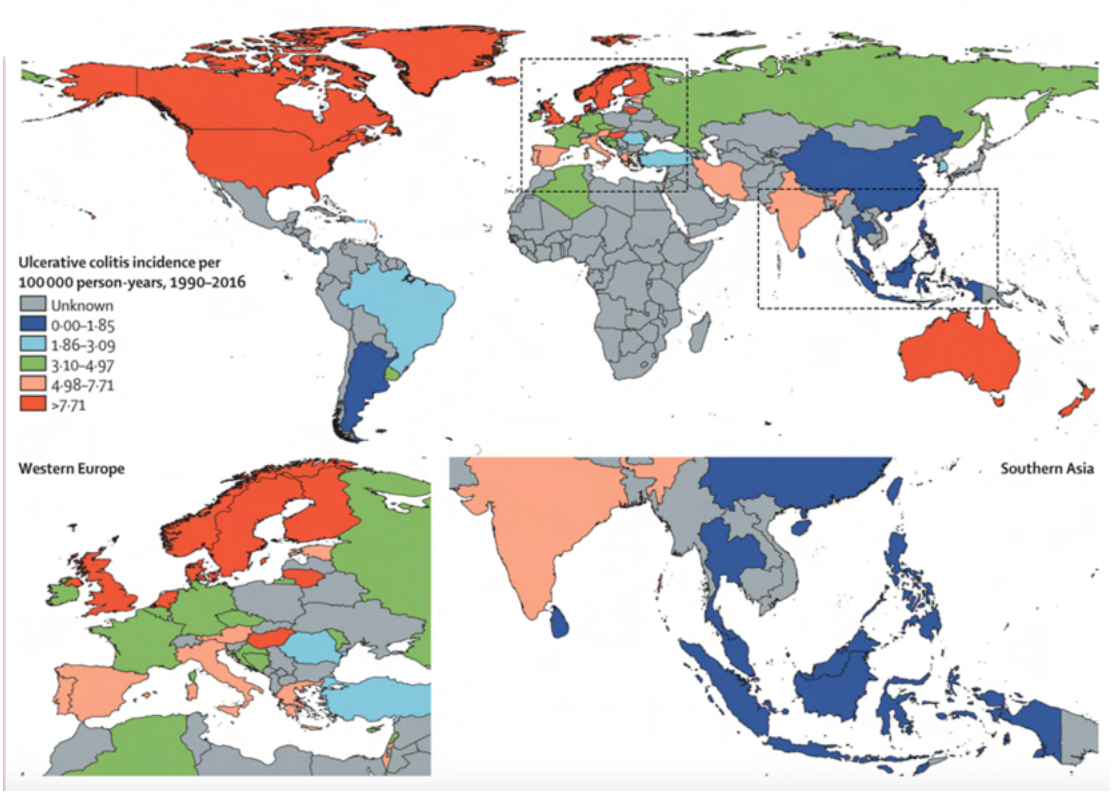


Figure 1.1 GEOPOLITICAL REPRESENTATION OF WORLDWIDE INCIDENCE OF UC EXPRESSED PER 100'000 PERSON-YEARS (Ng *et al*. 2020).

1.2 Clinical Presentations And Available Treatments

The clinical presentation of UC includes abdominal pain, increased frequency in bowel movement - often with blood in the stool (Baumgart and Sandborn, 2007; Ford, Moayyedi and Hanauer, 2013).

UC almost always involves rectum (proctitis) and progresses proximally in a continuous manner with variable extent of involvement of the colon. Early stage UC is characterized by inflamed, erythematous mucosa that bleeds easily (Figure 1.2 B). In severe cases extensive ulceration with pseudopolyps can be seen (Figure 1.2 D). UC involves only the mucosal and submucosal layers of the intestinal wall. Inflammation is

superficial and there is marked infiltration of neutrophils, macrophages and lymphocytes in the lamina propria layer. Neutrophils can intrude into the epithelial cell layer and cross into the intestinal lumen causing tissue damage resulting in loss of crypt architecture and goblet cell depletion (Gramlich and Petras, 2007; Fatahzadeh, 2009).

Fraction of UC patients will develop secondary condition outside gut named external manifestation, including inflammation in joints, skin problems and inflammation of different parts of eye (Levine and Burakoff 2011). Moreover, UC patients have increased risk of developing colorectal cancer when compared to the general population (Olén *et al*., 2020).

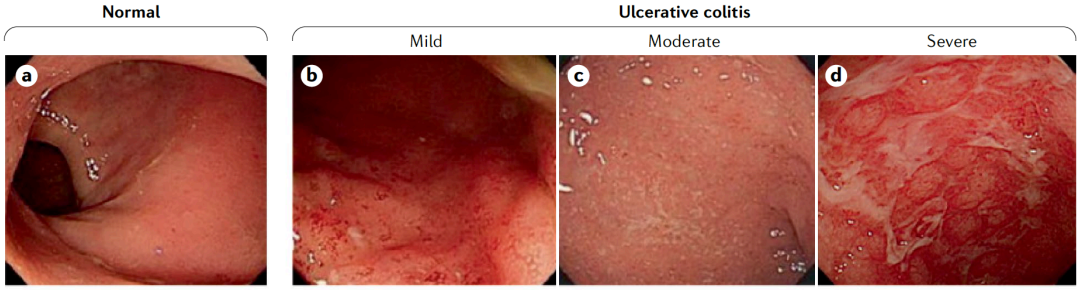


Figure 1.2 ENDOSCOPIC OUTLOOK OF B. MILDLY, C. MODERATELY AND D. SEVERELY AFFECTED INTESTINAL TISSUE OF UC PATIENTS IN COMPARISON TO A. HEALTHY SUBJECT (Kobayashi *et al.,* 2020)**.** *Endoscopy is irreplaceable in diagnosis and management of IBD. It plays a fundamental role in helping to distinguishing between UC and CD, yet, in some cases the histopathological and morphological features of UC and CD can overlap making clear diagnosis impossible.*

Currently there is no cure for UC. Instead therapy is focused on achieving clinal remission (Kobayashi *et al.,* 2020). UC management is depended on the disease severity and its evolution over time. For mild to moderate disease, 5 - aminosalicylate (5-ASA) are the first line of choice. If no improvement is seen, corticosteroids are added to the treatment (corticosteroids can be used only till the remission is achieved). If there is still no sign of improvement (thus the failure of conventional therapy) use of biologics, such as Infliximab, Adalimumab, Golimumab Vedolizumab and small molecule inhibitors, such as Tofacitinib, are tried (NICE guidelines 2019). However, there is risk that with time patient will either develop an immunological reaction or stop responding to biologic in use.

In acute severe UC, intravenous corticosteroids are tried first. However, in case of severe UC with no signs of improvement - surgical management is the only currently available lifesaving option left (NICE guidelines 2019; Kobayashi *et al.,* 2020).

1.3 Pathogenesis Of UC

Despite the major effort invested into understanding the disease cause, UC remains an idiopathic condition. Though, the exact aetiology of UC is still unknown, current hypothesis is that UC is result of uncontrolled immune response to environmental stimulus in genetically susceptible individuals (Figure 1.3) (Turpin *et al.,* 2018).

The main risk factor associated with IBD is a positive family history (Santos, Gomes and Torres, 2018). Having a first degree relative with CD or UC increases the risk of developing the same condition 8-fold and 4-fold, respectively (Moller *et al.*, 2015). This observation was supported by twin studies, where the concordance of IBD between monozygotic twins (20% - 55% in CD and 6.3% - 17% UC) was higher than concordance between the dizygotic twins (0% - 3.6% in CD and 0% - 6.3% UC) (Gordon *et al.*, 2015).

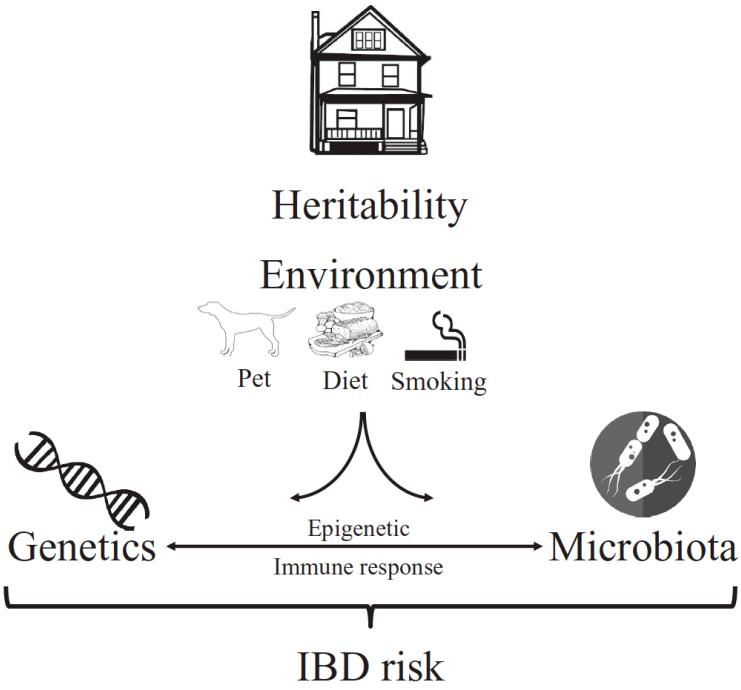


Figure 1.3 RISK FACTORS CURRENTLY BELIEVED TO HAVE SOME POSSIBLE IMPLICATIONS IN DEVELOPMENT OF IBD (Turpin *et al.,* 2018).

Indeed, genetic studies have identified more than 240 IBD risk associated loci, including risk variants falling into protein coding genes involved into innate and adaptive immunity, epithelial function and microbial clearance (Wellcome Trust Case Control Consortium, 2007; Anderson *et al.,* 2011; Jostins *et al.,* 2012; Liu *et al.,* 2015; de Lange *et al.,* 2017). However, individual effect size of associations discovered are modest and many healthy individuals carrying a risk associated variant never develop disease. In addition, in comparison to family studies, heritability explained by significant associations is much smaller. *Chen et al., 2014* showed that SNP-heritability accounts for 19% in comparison to 70% of pedigree heritability of UC (Chen *et al.,* 2014).

Since then, there have been multiple hypothesis, including polygenicity, gene-gene interaction, missing variants and questioning the adequacy of calculation itself, to explain the missing heritability (Génin 2019). However, immigration studies showing that children of immigrants from some of the low risk countries had the same incidence risk to develop IBD as children from non-immigrants (Li *et al.,* 2011; Benchimol *et al.,* 2015) and that the younger age of migrants themselves were associated with

the increased risk of developing the IBD (Benchimol *et al.,* 2015) suggests that the environmental exposure of possibly genetically predisposed individuals might be the key in IBD pathogenesis. Indeed, the presence of additional environmental triggers would help to explain the rapid increase in IBD epidemiology, particularly, the rise of IBD in newly developing countries. However, currently, there are many missing details concerning the most relevant environmental factors and their relative contributions to disease pathogenesis (Ananthakrishnan *et al.*, 2018).

1.4 Genetics Studies Of Common Disease

Early epidemiology studies showing strong family history of IBD paved the way to genetic studies desperately trying to identify causal variants. The first studies to identify a genetic locus predisposing to development of IBD used linkage disequilibrium analysis. These family pedigree studies assess the co-segregation of marker loci with the trait of interest (Ott, Kamatani and Lathrop, 2011). However, it soon become came clear that linkage disequilibrium analysis lack in power to identify causal genes for common diseases including UC (Altshuler, Daly and Lander, 2008). For immune mediated disease, including UC, CD, Rheumatoid Arthritis (RhArt) and Type II Diabetes (T2D), so-called "common disease / common variant" hypothesis suggests that multiple genetic mutations that occur with relatively high frequency in the general population may be associated with increased predisposition to disease development, but with each mutation having a relatively low relative risk (Kruglyak, 2008).

In order to identify such common but low penetrance variants, genotyping across multiple loci in large numbers of individuals is required. A major technical advance in this regard came with the development of large-scale genome-wide association studies (GWAS). These studies make use of simultaneous genotyping of multiple single nucleotide polymorphisms (SNPs), to compare DNA sequence variation between disease bearing individuals and the healthy population in a "hypothesis-free" manner that scans the entire genome for SNPs associated with increased risk of disease development (Wang *et al.*, 2005). Due to low rates of recombination between SNPs that are located close together across much of the genome, there is a strong probability for

groups of SNPs to be passed together through generations - a phenomena known as linkage disequilibrium (LD). Testing of relatively few SNPs (called tag SNPs) can therefore allow assessment of the association of much of the most common genetic variants across the whole genome, without the need for direct sequencing (Bush and Moore, 2012). Thus GWAS performed on large numbers of healthy and diseased individuals permit the identification of common genetic risk factors underlying complex genetic diseases or traits.

Currently there are more than 240 IBD risk associate loci identified (de Lange *et al.*, 2017), yet, the most aspects of disease pathogenesis remain unclear. It turned out that the main strength of GWAS design, which allowed to test from hundreds of regions simultaneously, become the Achilles heel of the study. GWAS allow the identification of risk regions centred on SNPs with the highest disease association signals (focal SNPs). However, these are typically in strong LD with other SNPs, thus, one cannot assume that focal SNP is a causal SNP without further study, including further characterization of potential functional roles for both the focal SNP and SNPs in LD (Schaub *et al.*, 2012).

1.5 Difficulties With GWAS Interpretation

Understanding and translating GWAS findings into causal regulatory pathways would provide an important stepping stone into development of better therapeutic strategies and diagnostics. However, despite the increasing attempts to dissect the causal variants and assign functional roles to UC risk variants and/or regions, progress towards understanding pathogenic mechanisms has been limited.

Fine-mapping is perhaps the most established computational approach used for refining evidence for a focal SNP in an LD block with multiple SNPs. Fine-mapping allow the assignment of probability of causality, where the ability to discriminate the causal variant depends on the effect of the variant and the sample size. The SNP showing the strongest probability is pronounced to be causal (Spain and Barrett, 2015). Dense genotyping in a further 67,852 individuals allowed pinpointing of 18 of 94 SNPs studied, with IBD risk statistically linked to a single causal variant. In an additional 27 cases,

SNPs were linked to a single variant with at least 50% probability of being causal (Huang *et al.*, 2017).

However, candidate variant identification alone does not typically provide biological insights into how the associated variant contributes to the studied phenotype. In rare instances, focal SNPs are associated with changes in the protein-coding sequence. For example, GWAS have demonstrated a non-synonymous mutation in Autophagy Related 16 Like 1 (ATG16L1) gene as a strong risk factor for the development of CD (Hampe et al., 2007). However, such instances of non-synonymous changes in protein-coding sequence are relatively rare (Jostins et al., 2012; Liu et al., 2015; de Lange et al., 2017a). Majority of UC associated risk loci are in the non-protein-coding part of genome.

IBD is not unique in having the majority of the underlying genetic risk described so far associated with non-protein-coding changes. In fact, 93% of GWAS disease- or trait-associated variants lie within non-protein-coding DNA regions and, thus, understanding the functional role of these non-protein-coding regions in general may help determine the potential biological role of disease or trait associated risk loci (Maurano *et al.*, 2012).

1.6 GWAS Risk Variants Enrichment On *Cis*-Regulatory Regions

In parallel to early genetics studies, large consortium named the ENCODE was formed with a goal to systematically map all functional elements in DNA. The ENCODE consortium defined functional elements as “discrete genomic segments that encode defined products or display a reproducible biochemical signature”. Together these data have transformed our understanding of the role of non-protein-coding DNA, by showing that 80.4% of the human genome has apparent biochemical functions and functional elements which are important in determining the cell identity (Ecker *et al.*, 2012). These therefore govern cell-type specific gene expression.

As part of the 2012 ENCODE consortium publication release, came the first detailed analysis of the human regulatory network (often divided into *cis*- and *trans*- regions

depending on their spatial locations). *Cis*-regulatory regions are DNA sequences (mainly non-protein-coding) involved in controlling transcriptional activity on the same chromosome and contain various *cis*-regulatory elements, including collections of binding sites for transcriptional factors (TFBSs) (Mathelier, Shi and Wasserman, 2015). Broadly, *cis*-regulatory elements may be recognized as acting in different functional categories. These include:

- Promoters - DNA regions where the pre-initiation complex (DNA-binding protein complex whose assembly governs mRNA transcription) binds (Riethoven, 2010);
- Enhancers - DNA segments that potentiate transcriptional activity of associated genes (Riethoven, 2010);
- Silencers - DNA segments that hinder expression of associated genes (Riethoven, 2010);
- Insulators - DNA regions that maintain discrete inter-domain boundaries when placed between enhancer-promoter or opened-closed chromatin (Riethoven, 2010);

Cis-regulatory regions, in particular enhancer regions, are enriched with GWAS risk variants (Ernst *et al.*, 2011; The ENCODE Project Consortium, 2012; Farh *et al.*, 2015; Kundaje *et al.*, 2015a). Intriguingly, variable DNA regions, such as SNPs, can modulate gene expression in a genotype specific manner (Dimas *et al.,* 2009; Fairfax *et al.,* 2012; Lee *et al.,* 2014). Moreover, variant can lead change in expression by modulating a function of cis-regulatory elements. Taken together, these observations could potentially explain and provide with mechanistical insights of how does UC risk associated variants located on non-protein-coding part of genome contributes to disease pathogenesis.

Indeed, experimental work by *Musunuru et al., 2010* showed that SNP located at the non-protein-coding region on 1p13 locus (previously identified as GWAS risk locus for Myocardial Infarction) alters the function of regulatory elements by creating a TFBSs which in turn increase the expression of *SORT1* gene. Follow up in mice gain-of-function studies showed that increase SORT1 expression correlates with decrease levels of

low-density lipoprotein cholesterol, which is well established risk factor for Myocardial Infarction (Musunuru *et al.,* 2010).

In summary, study combining GWAS data with predictions of regulatory element activity and gene expression data may dissect how risk variants located on non-protein-coding regions increase risk for UC development.

1.7 Determining Activity Of Regulatory Regions

One of the methods used to predict location and activity of potential regulatory regions utilizes chromatin accessibility landscape. *Thurman et al., 2012* showed that most TFBSs identified by the ENCODE Chip-seq were located within accessible chromatin regions (Thurman *et al.*, 2012).

Chromatin is the filamentary assembly of DNA and proteins in which only very short stretches of the naked helix are revealed. The fundamental unit of chromatin is the nucleosome which is made up from DNA twice wrapped around an octamer of core structural proteins known as histones (Lawrence, Daujat and Schneider, 2016). The topological distribution of nucleosomes across the genome is dynamic and reflective of chromatin functions. Chromatin accessibility is the extent by which DNA-binding factors are able to form a physical contact with the genomic region they would act on (Klemm, Shipony and Greenleaf 2019). Active promoters are associated with accessible chromatin, where binding of RNA polymerase II require rearrangement of chromatin structure (due to polymerase size) (Felsenfeld *et al.,* 1996). In addition, nucleosomes as well as other chromatin binding factors can provide steric hindrance to transcription factors (TF) binding and increase TF dissociation rates (Allis and Jenuwein 2016).

Chromatin accessibility can be assessed, at least in part, by mapping the sensitivity of chromosomal regions to enzymatic cleavage. An example of this approach involves assessing the sensitivity of isolated nuclear DNA to enzymatic cleavage by Deoxyribonuclease I (assay named - DNase seq) or Transposonase (assay named - Assay for Transposonase Accessible Chromatin Using Sequencing; ATAC seq).

ATAC seq allows fast and reliable assessment of the chromatin accessibility, nucleosome position and TF occupancy simultaneously. The underlying principle of ATAC seq is the use of an engineered hyperactive Tn5 transposase, which is loaded with adaptors for PCR amplification and sequencing. Exposure of nuclear DNA to this transposes under rate-limiting conditions therefore results in fragmentation of nuclear DNA and simultaneous tagging of the resulting DNA sequence in a manner that permits amplification and subsequent sequencing (Figure 1.4 A) (Buenrostro *et al.*, 2013, 2015). By integrating preferentially into open chromatin (steric hindrance in less accessible chromatin hinders access of the transposase), the engineered enzyme can thus be employed to allow for the amplification and preparation of sequencing libraries from small numbers of target cells previously not amenable to DHS methods (Figure 1.4 B). In addition, since TF binding to DNA sequence protects it from cleavage, ATAC Seq "footprints" may be analyzed for evidence of active TFBS (Buenrostro *et al.*, 2013).

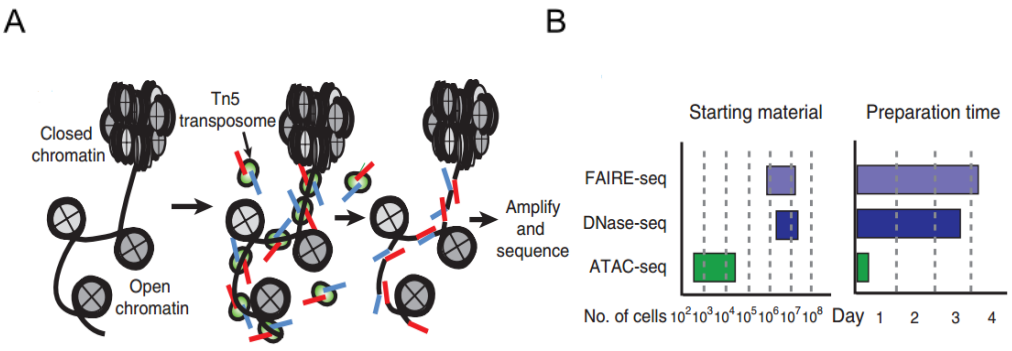


Figure 1.4 *A.* GRAPHICAL REPRESENTATION OF ATAC SEQ WORKING PRINCIPLES, *B.* TIME AND INPUT MATERIAL REQUIREMENTS DEPENDING ON SELECTED METHOD FOR OPEN-CHROMATIN ANALYSIS *(Buenrostro et al., 2013).*

ATAC seq has already successfully provided with meaningful insight into areas, such as cancer and developmental biology (Corces *et al.,* 2018, Berg *et al.,* 2019), neuronal biology (Fullard *et al.,* 2018) and complex disease, including, Systemic Lupus Erythematous (Scharer *et al.,* 2016) and Type 2 Diabetes (Bysani *et al.,* 2019). *Ludwig et al., 2019* employed ATAC seq to study the transcriptional factor dynamics through the erythropoiesis. When combined with GWAS data for erythroid cell treats, they showed that 19.4% of fine-mapped variants fell within accessible chromatin suggesting a potential for regulatory activity (Ludwig *et al.,* 2019).

1.8 Computational Models For Functional Interpretation Of GWAS Data

Possibly the greatest challenge in interpretation of GWAS data is the realization of endless complexity in transcriptional regulatory networks and functional diversity of different cell types. It is an important roadblock in a way of understanding the complex disease, such as UC, CD, T2D, RhArt, variety of cardiovascular and neurological conditions, where the pathogenesis goes beyond easily "studiable" Mendelian single-gene disorders. Instead, scientist is often faced with hundreds of risk regions, majority located in non-protein-coding part of genome and seemingly requiring one or more stimulus from an unknown environmental risk factors to initiate the disease development. Just the sheer quantity and the complexity of these data seem to exceed the capacity of what a conventional functional study could answer in any given amount of time (de Souza, Fiocchi and Iliopoulos 2017).

Computational analysis has proved itself to be capable of handling large data output genomic technologies returns. In the heart of computational analysis is a predictive model constructed by researcher, which usually is based on prior scientific observations (often created from large training data sets) and that could be applied other data sets (Brodland 2015).

Indeed, other functional genomics technologies, not just GWAS, is benefiting from computational biology. As an example, transcriptomics, proteomics, metabolomics often returns hundreds of hits. The common approach was to select only a handful of hits (often defined by p-value or log fold change) to follow up in functional studies. Whilst the functional studies are vital in validating the findings, selecting only a handful of hits is wasteful when the cost of experiment and clinical sample availability is considered. In addition, there is no guarantee that hits with the highest p-value or log fold change are relevant to treat studied. Computational analysis in turn maximizes the scientific output by allowing to iterate whole data set. In addition, single data layer, being it GWAS data or functional genomics data, on its own often fall short in providing full biological story. It has been recognized that combinatorial analysis of

multiple layers of -*omics* data have potential to further increase the functional understanding (Cano-Gamez and Trynka 2020; Nature Research Custom Media and Illumina 2020).

In summary, in complex disease space, predictive learning models could potentially unlock the genetic data penitential and translate these finding into the actionable knowledge, thus, understanding the functional consequence of these variants will increase the clinical translation. The few of most used computational models developed to assign the variant specific functionality are briefly discussed below.

1.8.1 GWAS SNP Enrichment Analysis

Currently there are number of different methods for enrichment calculation available, including ones aimed to dissect the disease specific cell types (SNPsea (Hu *et al.*, 2011), CHEERS (Soskic *et al.*, 2019)), propose the functional role of GWAS variants (GREGOR (Schmidt *et al.*, 2015)) and to partition the SNP heritability (LDSC Finucane *et al.,* 2015)). All with an underlying theme to deepen our understand regarding the possible functional role of risk variants.

The underlying assumption behind the enrichment analysis aimed to predict disease associated cell types and specific cell state, is that disease relevant cell types/states will have more pathogenicity/disease associated transcripts and genomic annotations and, hereby, they will be enriched for the disease associated variants. In this case genomic annotation refers to the functional genomic elements that has a regulatory activity (Cano-Gamez and Trynka 2020).

Farh *et al* showed that most of risk variants for autoimmune disease, including CD and UC, are enriched in enhancers and promoters active in $CD4^+$ T cells subpopulations. Notably, causal variants associated with UC were also increased in *cis*-regulatory elements from colonic mucosa (Farh *et al.*, 2015).

Enrichment methods looking to understand the functional role of risk variants, work under similar assumption as ones used to prioritize the disease related cell types, thus

if disease associated risk variants act via altering the function of regulatory regions, then they should fall within the corresponding genomic annotations.

Maurano et al., 2012 mapped chromatin accessibility (DNase seq) in 349 cell-types and showed that GWAS identified disease associated variants were enriched in in accessible chromatin sites, whereas proportion of the GWAS SNPs overlap with DHS in disease relevant cell types, thus showing cell type contribution to phenotype (Maurano *et al.*, 2012). Altogether supporting the hypothesis that GWAS identified disease associated variants act by altering the function of the non-protein-coding regions in a cell type specific manner.

1.8.2 Colocalization Analysis

Variable DNA regions, such as SNPs, can modulate gene expression in a genotype specific manner (Albert and Kruglyak, 2015). Following this observation, presence of the SNP should be reflected within expression of target gene as well as the functional genomic elements which activity variant affects. Statistical model used to link the genotype with the functional data (expression, exon splicing, chromatin accessibility and marks) is called quantitative trait locus (QTL) analysis (Miles and Wayne 2008). For this approach, no prior knowledge of the functional mechanisms is required (Albert and Kruglyak, 2015). However, large numbers of donors are required.

Colocalization analysis aims to integrate the GWAS data with QTL data and identify if the same variant is causal in both GWAS and QTL studies (Wen, Pique-Regi and Luca 2017). The analysis is markedly complicated by the LD and possibility of multiple causal variants in loci (Cano-Gamez and Trynka 2020). Moreover, variant can have different functional roles in different cell types or in the same cell type under different stimulus (Gerrits *et al.*, 2009; Fairfax *et al.,* 2012). Therefore, it is important to note that colocalization analysis cannot establish causality, but can be used for hypothesis generation.

Chun et al., 2017 looked to determine if autoimmune and inflammatory diseases, including Multiple sclerosis (MS), IBD, UC, CD, Type 1 diabetes (T1D), Celiac disease and

RhArt, risk associated variants colocalized with *cis*-expression QTLs identified in $CD4^+$ T cell, $CD14^+$ monocytes and lymphoblastoid cell lines (Chun *et al*., 2017). Together they showed that only minority of the risk associated loci colocalized with eQTLs and concluded that proximity between GWAS risk variants and QTLs is not enough to propose the function of variant. However, the same as other studies, they showed that fraction of overlaps observed were cell type specific.

Currently majority of computational models created to propose functional role of GWAS data integrates risk variants with only a single layer of functional data. This approach has been proven to be great for gaining some insight into variant functionality. However, it does not provide with full mechanistic insight of how risk variant could lead to disease pathogenies. Multi-dimensional data sets that combine multiple elements are required in order to correlate the nucleotide variations with change in nearby regulatory functions and gene expression.

MOLOC is a colocalization analysis aimed to integrate GWAS summary statistics with multiple QTLs from functional -*omics* data (Giambartolomei *et al.,* 2018). Authors used expression QTL and methylation QTL to identify regulatory effects of Schizophrenia risk loci. MOLOC does not detect causal relationships among the associated traits, yet they showed that addition of an extra layer of functional data increased the gene discovery by 1.5 times. However, the interpretability got more complex than from pairwise comparison, with single locus having 15 different possible hypotheses.

In summary, the integration of GWAS data with functional genomics data has been successful for new hypothesis generation and translation of genetic findings into something easier to follow up. However, integration of pairwise dataset alone has proven to be extremely challenging and the difficulty is only increasing with more layer added. Hereby, conventional functional studies in vital to fully validate the findings from the computational analysis.

1.9 Disease Specific Cell Type(s) And Their Rolle In GWAS Interpretation

The importance of understanding non-coding DNA changes within the context of disease relevant cell types is highlighted by the fact that single variant can have a different (even opposing functions) in different cell types (Dimas *et al.,* 2009, Fairfax *et al.,* 2012). It is mirrored by the cell type specific action of regulatory elements, particularly enhancers. Ernst *et al* 2011 showed that enhancer activity is correlated with regulation of tissue dependent gene expression (Ernst *et al.*, 2011). Thus, cell type specific studies are vital for correct interpretation of regulatory element function and understanding how risk loci alter their function and lead to increased risk for disease development.

However, to date most attempts to annotate UC risk associated GWAS data have relied upon data from peripheral blood, tissue and *ex vivo* cultured primary cell lines from healthy individuals (Kabakchiev and Silverberg, 2013; Singh *et al.*, 2015; Peloquin *et al.*, 2016; Momozawa *et al.*, 2018). Considering the above, we identified a need for functional genomics data in UC relevant cell types.

In time of experimental design, it was unclear which cell types are UC relevant, therefore we reasoned that due to the inflammatory nature of the disease and due to its intestinal location, purified cell populations from intestine would be the most attractive target to use in our study. In addition, previous data from our lab showed that GWAS risk loci for CD and UC where enriched in genes upregulated in intestinal lymphocytes when compared to T cells from peripheral blood (Raine *et al.,* 2015). Further highlighting the possibility of intestinal immune cells being disease specific. In addition, by looking at that time available literature we reasoned that intestinal epithelium (due genetic studied linking epithelial malfunction with IBD) and blood lymphocytes (due enrichment studies showing IBD risk associated variant enrichment in $CD4^+$ T cells) could be an interesting and possibly disease relevant cell populations as well. The intestinal cell types used in this study are briefly discussed below.

1.9.1 Intestinal Epithelium

Both large and small bowel are in constant exposure to luminal contents and tight regulation is necessary to maintain the symbiotic relationship between the external environment and self. The intestinal epithelium, coated in layers of mucus, provides the most basic physical barrier between the microbiome and underlining lymphoid tissue (Chelakkot, Ghim and Ryu, 2018). (The microbiome is term given for all micro-organisms, such as, bacteria, fungus, protozoa and viruses that inhibits specific environment, in this case the human gastrointestinal tract (Barko *et al.*, 2018)).

The epithelium is in constant dynamic biochemical communication with the microbiome and other cells located within the mucosa, and thus contributes to immunomodulatory, metabolic and digestive functions (Okumura and Takeda, 2017; Allaire *et al.*, 2018). Intestinal epithelia cells express key receptors of the innate immune system: Toll like receptors (TLRs) and nucleotide binding oligomerization domain-like receptors (NODs). These direct the cells to elicit both pro- and anti- inflammatory signals. TLR and NOD belongs to the pattern-recognition receptors (PRRs) family and recognize foreign bacterial, viral and fungal structures in highly context specific manners (Peterson and Artis, 2014).

Any impairment in epithelial cell function could potentially lead to disruption in homeostasis and result in inflammation. This has been supported by various murine models where disruption of both – the physical barrier and/or epithelial function - led to spontaneous inflammation (Nenci *et al.*, 2007; Johansson *et al.*, 2008; Wirtz *et al.*, 2017). Mucin 2 (MUC2) is one of the main mucins of colonic mucus. Mice with MUC2 -/- deletion suffer from increased contact with intestinal microbiota which leads to spontaneous colitis by 7 weeks of age (Johansson *et al.*, 2008).

1.9.2 Intraepithelial Lymphocytes

Above the basement membrane, interspersed between epithelial cells, is a specialized subclass of T lymphocytes, known as intraepithelial lymphocytes (IEL) (Guy-Grand, Griscelli and Vassali, 1974). Although a range of immunocytes may be present in this

layer, the dominant and best characterized population is intra-epithelial T cells and in further text we use IEL to denote these epithelial resident T cells only. Despite their early discovery and extensive studies, understanding of the role of human intestinal IELs is still limited with most information arising from studies in rodents. Intestinal IELs are heterogeneous and the number and subtype profile of IELs vary between different parts of intestine (Mowat and Agace, 2014; Mayassi and Jabri, 2018).

1.9.3 Lamina Propria Residing T Cells

Below the intraepithelial layer lies the lamina propria layer. This is bounded by the thin basement membrane of epithelial cells above and the muscularis mucosae below and is home for variety of immune cells including LPL (Reed and Wickham, 2009). These LPLs are numerically more abundant that IELs and appear to possess distinct biology.

In humans, naïve T lymphocytes travel to distal regions, where after antigen priming they develop into T effector cells (T_E), from which subpopulation persists as T memory cells (T_M) (Kumar, Connors and Farber, 2018). The fundamental benefit of immunological memory is a faster response upon reencountering of an antigen. T_M can be further divided into smaller functionally and phenotypically distinct subsets. The heterogeneity of human T_M has been defined based on differential expression of costimulatory and adhesion molecules, such as, CD45 isoforms (CD45RO, CD45RA) and lymph node–homing receptors Cluster of differentiation 62 L and CC-chemokine receptor 7 (CD62L and CCR7) (Sathaliyawala *et al.*, 2013).

Thome et al 2016 showed that both ileum and colon of young adults (15 – 25 years of age) are predominantly occupied by T effector memory cells (T_{EM}) whereas the rest of T_M subsets, such as, T central memory (T_{CM}) and T terminal effector memory (T_{EMRA}) are almost absent (Thome *et al.*, 2016). However, recent studies in mice have led to the identification of yet another T_M subset, which does not recirculate and, thus, is named tissue resident memory T cells (T_{TRM}) (Klonowski *et al.*, 2004; Masopust *et al.*, 2010). It is not yet clear if the same populations persist in human, but *ex vivo* studies of human T cells have showed that non-lymphoid tissue residing T_{EM} express Cluster

of differentiation 69 (CD69), a marker also expressed by majority of murine T_{TRM} (Sathaliyawala *et al.*, 2013). CD69 is an early T cell activation marker which can block T cell exit from tissue by suppressing the sphingosine 1-phosphate receptor type 1 (S1PR1), a molecule required for cell emigration (Sheridan and Lefrançois, 2011). *Kumar et al 2017* showed that a considerable majority (<90%) of intestinal T_{EM} cells express CD69, which is not expressed by the T_{EM} located in blood (Kumar *et al.*, 2017).

In common with type a IELs, lamina propria T cells and B cells are associated with a "classical" immune response, undergoing priming within in secondary lymphoid tissue. Typically, evading pathogens breaching the mucosal surface are captured by Antigen Presenting Cells (APCs), which then migrate to secondary lymphoid tissue and induce T cell activation and imprinting (Iwasaki and Kelsall, 2001; Mowat *et al.*, 2003). Naïve $CD8^+$ T cells, on activation, differentiate into $CD8^+$ T cytotoxic T cells, whereas naïve $CD4^+$ T cells, differentiate into distinct T helper subsets based on the stimulatory cytokines (Szabo *et al.*, 2000; Mullen *et al.*, 2001; Fields, Kim and Flavell, 2002; Mangan *et al.*, 2006; Veldhoen *et al.*, 2008). Once priming and imprinting has occurred effector T cell enter efferent lymphatics and subsequently return to bloodstream (Girard, Moussion and Förster, 2012). In the circulation, T cell homing molecules interact with adhesion ligands expressed by endothelium and hold lymphocytes in place thus facilitating their extravasation into intestinal tissue (Berlin *et al.*, 1993; Lefrançois *et al.*, 1999). However, in light of the recent discovery that most LPL resident T cells belong to the T_{TRM} population, which do not recirculate, it is yet to be determined if these T cells can be reactivated in lymphoid tissue independent manner, and what the biological consequences might be.

1.9.4 Lamina Propria Residing B Cells

Lamina propria B cells differentiate into plasma cells and together with a transmembrane epithelial glycoprotein receptor known as the polymeric Ig receptor (pIgR) maintain luminal immunoglobulin A (IgA) and M (IgM) levels. IgA is the predominant immunoglobulin produced in intestinal tissue, is protected from protease cleavage and prevents antigen encounter with the epithelial cell surface and, thus, contributes

to immune exclusion (Brandtzaeg *et al.*, 1999). In human, IgA is dimeric and consist of both IgA1 and IgA2 subclass. The ratio of IgA1 to IgA2 changes along the bowel, where IgA1 is preferentially produced by plasma cells in the small intestine. In colon the ratio is almost equal. In addition to pIgR, enterocytes also express the transferrin receptor Cluster of differentiation 71 (CD71) which can bind to IgA1 and facilitate luminal antigen uptake (Spencer and Sollid, 2016).

2. Aims

We hypothesize that UC risk variants can alter the function of regulatory elements in disease relevant cell-types and thus contribute to disease pathogenesis. We speculated that by combining GWAS data with transcriptomics data and cis-regulatory element activity (measured by chromatin accessibility) we may be able to understand the functional role of UC associated risk variants.

The specific goals for this project were:

1) To assess performance of commercially available anti-GPR15 antibodies. (Side project)

2) To compare the chromatin state and transcriptional activity between healthy volunteers and UC patients at purified single cell population level.

3) To combine functional genomics and/or GWAS data to assess:

- How much of the chromatin conformational changes are reflected by gene expression and vice versa.
- GWAS immune disease and treat associated loci enrichment with differentially accessible chromatin regions and differentially expressed genes.
- How GWAS risk variants could potentially alter the chromatin state and genes expression.

3. Materials And Methods

Human Biological Samples were sourced ethically and their research use was in accord with the principles of the informed consent.

3.1 Recruitment Of Healthy Donors And UC Patients

The main study had a two stages of participant recruitment: at the first stage samples were collected for both ATAC Seq and RNA Seq analysis, whereas later recruitment was purely to increase sample numbers for the RNA Seq study. All samples for this study were collected at the Addenbrookes Hospital – Endoscopy Unity.

A total of 39 individuals, undergoing medically indicated colonoscopy, were recruited for biopsy and blood collection as part of this study. Eight pinch biopsies were taken from the SC and TI respectively. Biopsies were collected into the RPMI medium 1640 (Life Technologies, 11875-093) on ice, along with 5-10ml blood (into tubes containing 15mg EDTA). Only biopsies taken from SC were used in this PhD project. All TI biopsies were processed till sequencing library stage and frozen in -80^{0}C.

Prospective volunteers (UC cohort) were approached if they either had previously been diagnosed with UC (identified from health records stored on hospitals electronic data base) or if their symptoms during the time of endoscopy were ones of the UC. Control samples were obtained from healthy subjects undergoing colonoscopy for screening purposes. Donor fit for this study was re-evaluated after the results from the gastrointestinal endoscopy and histological examination were added on their health record. The main exclusion criteria for the study were old age and presence of other immune mediated disease.

After careful review of medical history and sequencing quality metrics, 29 participants (15 UC patients and 14 healthy controls, Table 3.1) passed all study requirement and, thus, were used for downstream analysis. The final UC cohort was heterogeneous in their disease history, in particular with regard to disease duration and severity. To control some of this variation, for further analysis, UC patients were split into two subgroups - UC with microscopically and/or macroscopically inflamed SC (UC_i) or UC without inflammation in the SC (UC_n).

3.2 Volunteer Demographics

Additional factors that could potentially contribute to the inter-individual variation were also recorded. These included patient age, sex, smoking history, other conditions diagnosed by time of sample collection and medication (Table 3.1). As part of this study we did not collect any supplementary clinical information.

Since only a subset of the initially collected samples were used in the final analysis, as shown in Table 3.1, the mean age ± standard deviation and gender ratio was calculated for each population individually, based upon only those samples that contributed to the final analysis (Table 3.2). The mean age of controls ranged from 47.7 – 57.6 years, whereas for UC_i it was from 35.5 – 42.5 years. The oldest population was UC_n, where mean age varied from 59.2 -64.3 years. Due to low sample numbers in each category it was not possible to conduct reliable significance estimations. Just by looking at crude age ranges, it seems that each individual condition is very well matched in the inter-individual age. Moreover, UC_i falls into the expected age for disease onset in adults. However, for very heterogenous disease, such as IBD, study evaluation based on age alone might be very mis-informative. Hereby, to understand how age affect the data, it should be evaluated in light of accompanying factors, including the disease onset, duration, severity and remission-relapse.

With respect to gender, there was a marked female dominance seen in all the groups. The sex effect on chromatin accessibility and ncRNA has been reported (Qu *et al.*, 2015). Though we excluded both sex chromosomes, we could not completely eradicate any possible gender disbalance in some of the comparisons.

Table 3.1 PARTICIPANT DEMOGRAPHICS AND SAMPLES USED IN THE FINAL ANALYSIS. *Fields colored in green shows samples that were included in the final analysis, whereas in red shows samples that had either failed the QC (including the threshold for cell counts). Ex - Ex-smoker; F – Female; IEL – Intraepithelial lymphocytes; LPL – Lamina Propria; M – Male; MF – Monocytes and macrophages; N/A – Status unknown; N - Non-Smoker; S – Smoker; T_{EM} – T effector memory; UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon; UC(N) - Ulcerative colitis patient with non-inflamed Sigmoid colon.*

Participant No	Condition	Secondary Condition	Age	Sex	Smoking History	Year of Recruitment	Current Medications	RNA-seq				ATAC-Seq									
								Blood		Sigmoid Colon		Blood				Sigmoid Colon					
								$CD4^+$ T_{EM}	$CD19^+$ B cells	LPL $CD4^+$ T_{EM}	LPL $CD19^+$ B cells	$CD8^+$ T_{EM}	$CD4^+$ T_{EM}	$CD14^+$ MF	$CD19^+$ B cells	Epi-thelium	IEL $CD8^+$ T_{EM}	IEL $CD4^+$ T_{EM}	LPL $CD8^+$ T_{EM}	LPL $CD4^+$ T_{EM}	LPL $CD19^+$ B cells
1	Control		79	F	N	2015	Aspirin														
2	Control		61	F	S	2015	Citalopram, Lisinopril														
3	Control		63	M	EX	2017	Losartan, Bisoprolol, Atorvastatin, Aspirin, Allopurinol														
4	Control	IBS	49	F	N	2015	Levothyroxine, Inhalers (Asthma)														
5	Control	Arthritis	67	F	EX	2015	Levothyroxine, Hydroxychloroquine, Sulfasalazine														
6	Control		28	M	N	2017	No Meds														
7	Control		46	M	N/A	2015	No Meds														
8	Control		35	F	N	2017	No Meds														
9	Control		52	M	N	2017	No Meds														
10	Control		44	F	S	2015	No Meds														
11	Control	Sjögren's	52	F	N	2015	No Meds														
12	Control		32	F	EX	2015	No Meds														
13	Control		18	M	N	2015	No Meds														
14	Control	IBS	52	F	N	2015	No Meds														
15	UC(I)		25	M	N	2017	IV Steroids, Hydrocortisone														
16	UC(I)		33	M	N	2015	No Meds														
17	UC(I)		49	F	N	2017	Pentasa, Vitamin B12, Mirena Coil														
18	UC(I)		45	F	N	2015	No Meds														
19	UC(I)		36	M	EX	2015	Mesalazine														
20	UC(I)		40	F	EX	2015	No Meds														
21	UC(I)		20	M	N	2016	No Meds														
22	UC(I)	Coeliac	33	F	N	2016	Pentasa, Salofalk														
23	UC(I)		39	M	EX	2016	Pentasa														
24	UC(I)		56	M	N	2016	Mesavancol, Budenofalk Rectal Foam														
25	UC(N)		46	F	N	2015	No Meds														
26	UC(N)		57	F	N	2017	Asacol, Prednisolone, Adcal D3, Asacol Suppositories														
27	UC(N)		63	F	N	2015	Azathioprine, Sulfasalazine														
28	UC(N)		72	F	N	2015	Pentasa														
29	UC(N)		58	F	N	2015	Pentasa, Mesalazine														

Table 3.2 AGE, SEX AND SAMPLE NUMBERS AND RATIOS FOR EACH INDIVIDUAL POPULATION USED FOR DIFFERENTIAL ACCESSIBILITY OR DIFFERENTIAL EXPRESSION ANALYSIS.

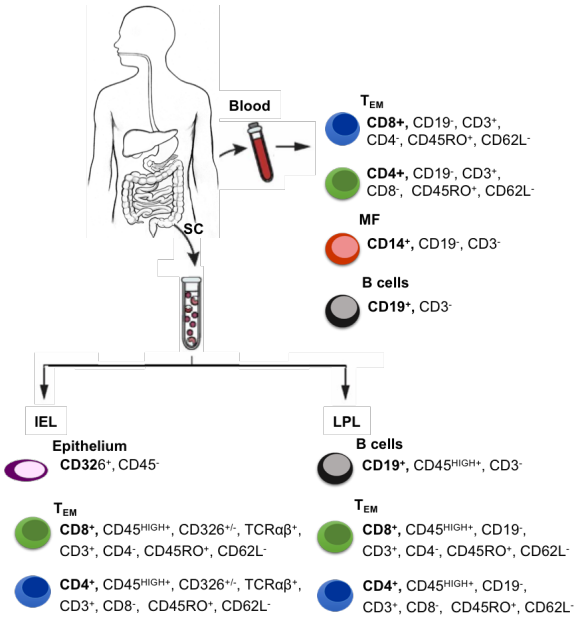
	RNA-seq				ATAC-seq									
	Blood		Sigmoid Colon		Blood				Sigmoid Colon					
	$CD4^+$ T_{EM}	$CD19^+$ B cells	LPL $CD4^+$ T_{EM}	LPL $CD19^+$ B cells	$CD8^+$ T_{EM}	$CD4^+$ T_{EM}	$CD14^+$ MF	$CD19^+$ B cells	Epithelium	IEL $CD8^+$ T_{EM}	IEL $CD4^+$ T_{EM}	LPL $CD8^+$ T_{EM}	LPL $CD4^+$ T_{EM}	LPL $CD19^+$ B cells
Total Number of Samples	22	26	20	26	15	15	17	14	14	11	10	16	14	15
Sample Ratio (C:I:N)	10:9:3	12:9:5	8:8:4	13:8:5	7:4:4	6:5:4	8:5:4	6:4:4	9:2:3	3:4:4	3:3:4	8:4:4	5:5:4	7:5:3
Sex Ratio														
Female (C:I:N)	6:3:3	8:3:5	5:4:4	8:3:5	6:2:4	5:3:4	7:3:4	5:2:4	7:2:3	3:2:4	3:3:4	8:2:4	4:3:4	5:3:3
Male (C:I:N)	4:6:0	4:6:0	3:4:0	5:5:0	1:2:0	1:2:0	1:2:0	1:2:0	2:0:0	0:2:0	0:0:0	0:2:0	1:2:0	2:2:0
Mean Age ± SD														
C	50.30 ± 6.18	50.70 ± 15.1	50.88 ± 13.58	48.15 ± 17.06	48.71 ± 20.44	50.00 ± 22.3	50.25 ± 19.42	47.67 ± 21.53	49.78 ± 18.22	57.33 ± 18.93	53.33 ± 23.8	54.50 ± 14.43	57.60 ± 14.98	50.71 ± 19.18
I	37.33 ± 11.39	37.33 ± 11.39	35.50 ± 9.77	36.88 ± 11.85	36.75 ± 5.68	37.40 ± 5.13	37.40 ± 5.13	36.75 ± 5.68	42.50 ± 3.54	36.75 ± 5.68	39.33 ± 6.03	38.50 ± 5.20	37.40 ± 5.13	37.40 ± 5.13
N	62.33 ± 8.39	59.20 ± 9.47	59.50 ± 10.91	59.20 ± 9.47	59.75 ± 10.84	59.75± 10.84	59.75 ± 10.84	59.75 ± 10.84	64.33 ± 7.09	59.75 ± 10.84	59.75 ± 10.84	59.75 ± 10.84	59.75 ± 10.84	64.33 ± 7.09

3.3 Purification Of Individual Cell Types

Cells were first purified from peripheral blood and pinch biopsies following the protocol already established in lab and detailed below.

Flow cytometry was selected for final cell type purification. First, all identified markers were tested, multicolor panels and sorting strategy developed and optimized (Figure 3.1).

Figure 3.1 SUMMARY OF CELL POPULATIONS AND THEIR SORTING MARKERS. *IEL – Intraepithelial lymphocytes; LPL – Lamina Propria; MF – Monocytes and macrophages; T_{EM} – T effector memory.*



3.3.1 Peripheral Blood Mononuclear Cell Isolation

Human blood was diluted 1:1 in Dulbecco's Modified Eagle's medium (DMEM) GlutaMAX (Life Technologies, 31966-021) + 0.5% Bovine Serum Albumin (BSA, Sigma A3912). Lymphoprep (Axis-Shield) was used for density-based separation (800xg, 20min, 4 ºC). Interface was collected in 15ml Falcon tube and washed with 10ml DMEM + 0.5% BSA, twice (250xg, 5min, 4 ºC). Peripheral blood mononuclear cell (PBMCs) were stained with an antibody panel designed.

3.3.2 Lamina Propria And Intraepithelial Cell Separation

Biopsies were incubated in calcium and magnesium free Hank's Balanced Salt Solution (HBSS, Life Technologies 14170088) with 2mM EDTA (Sigma E5134) and 0.5% BSA. Samples were agitated at 900rpm for 32 min (20 min on shaking platform + 12 min by vortex) at room temperature. 2 x 6.5 µl 1M DL-Dithiotheritol (DTT) was added at 0 min and 10 min respectively. Media was changed at 24 min, 28 min and 32 min, and the supernatants containing the IEL fraction were filtered via 70µm cell strainer (Falcon 352350). The flow through containing IEL was spun down at 250xg for 7min at 4^{o}C and resuspended in 200µl HBSS + 0.5% BSA + 2mM EDTA and left on ice.

Remaining tissue patches were washed with 10ml DMEM + 0.5% BSA, resuspended in 1ml of the same media and incubated for 1.5h at 37^{o} C after addition of 10µl type IV collagenase (at 128U/ml, Sigma C1889) and 10µl type V DNAse II (10U/ml, Sigma D8764). Tissue fragments were then further fragmented via grinding through a 70µm cell strainer. The flow through containing LPL was spun down at 250xg for 6min at 4^{o} C and resuspended in 200µl DMEM + 0.5% BSA. IEL and LPL were stained with anti-body panel designed.

3.3.3 Cell Staining

Cells were incubated with an antibody cocktail (final volume = 100µl) for 20 min on ice in dark. Cells were then washed by adding 2ml-3ml media and spun at 250xg for 4min, pipetted through 70µm cell-strainer and resuspended in 200µl fresh, cell type com-patible, media for Fluorescence activated cell sorting (FACS) or FACS Analysis. All an-tibodies used in study, including test runs, are showed in Table 3.3. Unfortunately, the viability staining was not included in our experiments.

Table 3.3 LIST OF ALL ANTIBODIES USED DURING THE SEQUENCING STUDY. *Table show all markers used in either sorting strategy development or in actual study and the distributer, catalogue number, clone, fluorochrome and concentration used for staining lamina propria, intraepithelial and peripheral blood mononuclear cell suspensions. LP - Lamina propria; IEL - Intraepithelial lymphocytes; Cat. No. - Catalogue Number.*

Marker	Clone	Fluorochrome	LP Conc.	IEL Conc.	Blood Conc.	Company	Cat. No.
CD14	M5E2	BV650	-	-	2:100	BioLegend	301835
	M5E2	PE	-	-	2:100	BioLegend	301805
CD19	HIB19	PB	2:100	-	-	BioLegend	302224
	HIB19	FITC	3:100	-	2:100	BioLegend	302206
CD3	SK7	PerCP	1:100	1:100	1:100	BioLegend	344813
	SK7	PECY7	1:100	1:100	1:100	BioLegend	344815
	UCHT1	BV421	1:100	1:100	1:100	BioLegend	300433
CD326	9C4	FITC	-	3:100	-	BioLegend	324204
	9C4	PB	-	2:100	-	BioLegend	324218
	9C4	BV650	-	2:100	-	BioLegend	324225
CD4	RPA-T4	BV570	1:100 & 2:100	1:100 & 2:100	1:100 & 2:100	BioLegend	300534
	OKT4	BV605	2:100	2:100	2:100	BioLegend	317437
CD45	HI30	PECY7	1:100	1:100	-	BioLegend	304015
CD45RO	UCHL1	BV785	2:100	2:100	2:100	BioLegend	304233
	DREG-56	APCeF780	2:100	2:100	2:100	e-Bioscience	47-0629-42
CD8a	HIT8a	AF700	1:100	1:100	1:100	BioLegend	300920
TCRαβ	IP26	BV421	-	3:100	-	BioLegend	306721
	IP26	PECY5	-	3:100	-	BioLegend	306710
	IP26	PE	-	3:100	-	BioLegend	306707
CD62L	DREG-56	APC	2:100	2:100	2:100	BioLegend	304810
	DREG-56	APCeF780	2:100	2:100	2:100	e-Bioscience	47-0629-42

3.4 Flow Cytometry

3.4.1 Cell Phenotyping

We used BD Fortessa and BD Fortessa 2 cell analyzers to develop a sorting panel and strategy for both RNA seq and ATAC seq experiments. Instruments were located at the NIHR Cambridge BRC Cell Phenotyping Hub.

For our anti-GPR15 antibody validation experiments we used FACS Canto and BD Fortessa, located at the MedImmune, Granta Park.

For data analysis FlowJo_VX (FlowJo LLC, Ashland) software were used.

3.4.2 Cell Sorting

For both - sequencing and anti-GPR15 antibody validation experiments, BD Aria III – Fusion and BD Aria III cell sorters were used. Sorting was performed by the NIHR Cambridge BRC Cell Phenotyping Hub.

First, populations of size of 1000-5000 events intended for ATAC Seq were sorted into 350µl DMEM + 0.5% BSA. Then, if there were enough leftovers for a given sample, same populations were sorted for transcriptomics study. Cells for RNA Seq were sorted into either 350µl of RLT Buffer (Qiagen, 79216) or RLT plus Buffer (Qiagen, 1053393). The RLT Buffer was used for all phase 1 recruitment samples, where RLT plus Buffer for all latter samples. Figure 3.2 shows the FACS gating strategy for Blood, IEL and LPL samples, where each dot is representing a single cell.

A

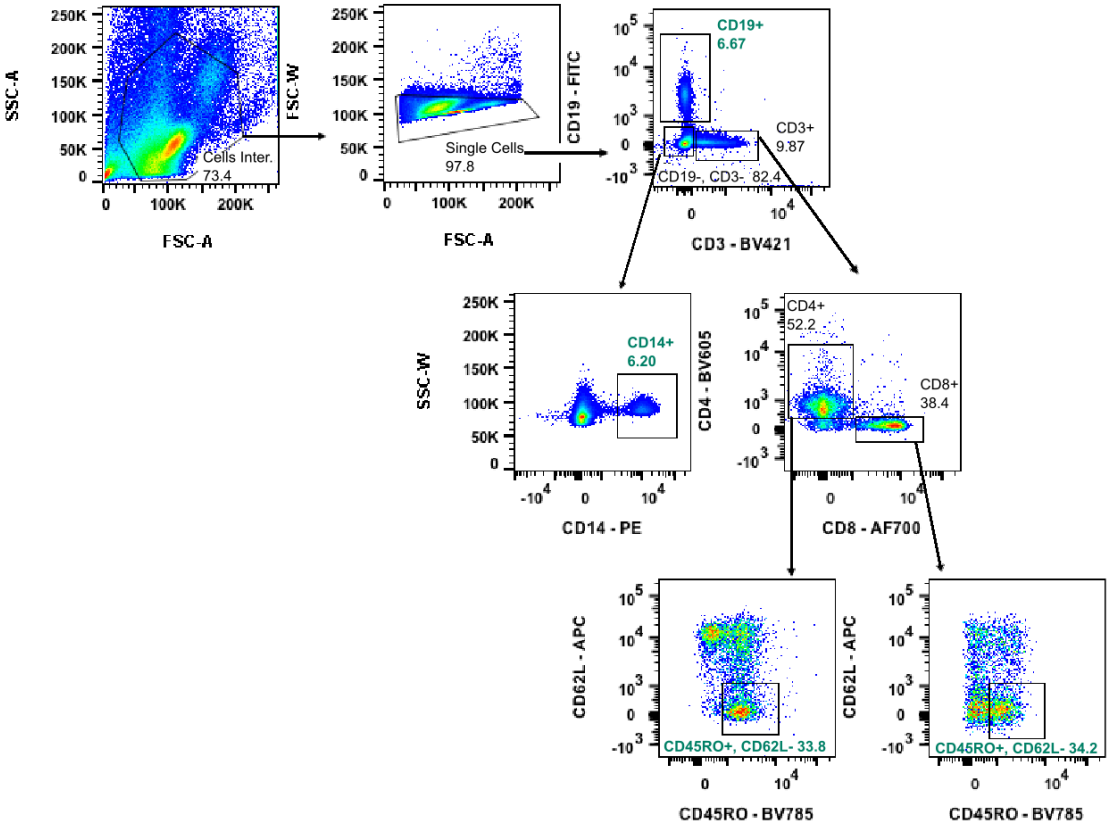
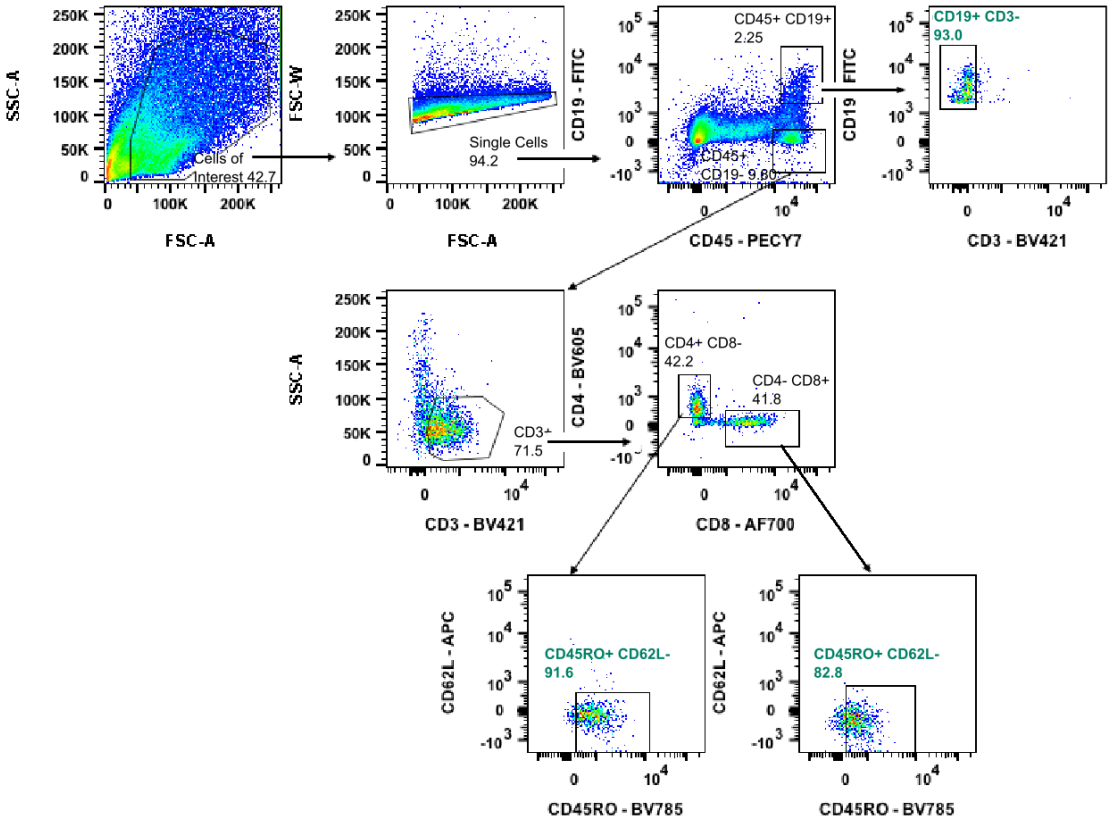


FIGURE CONTINUED IN NEXT PAGE

B



C

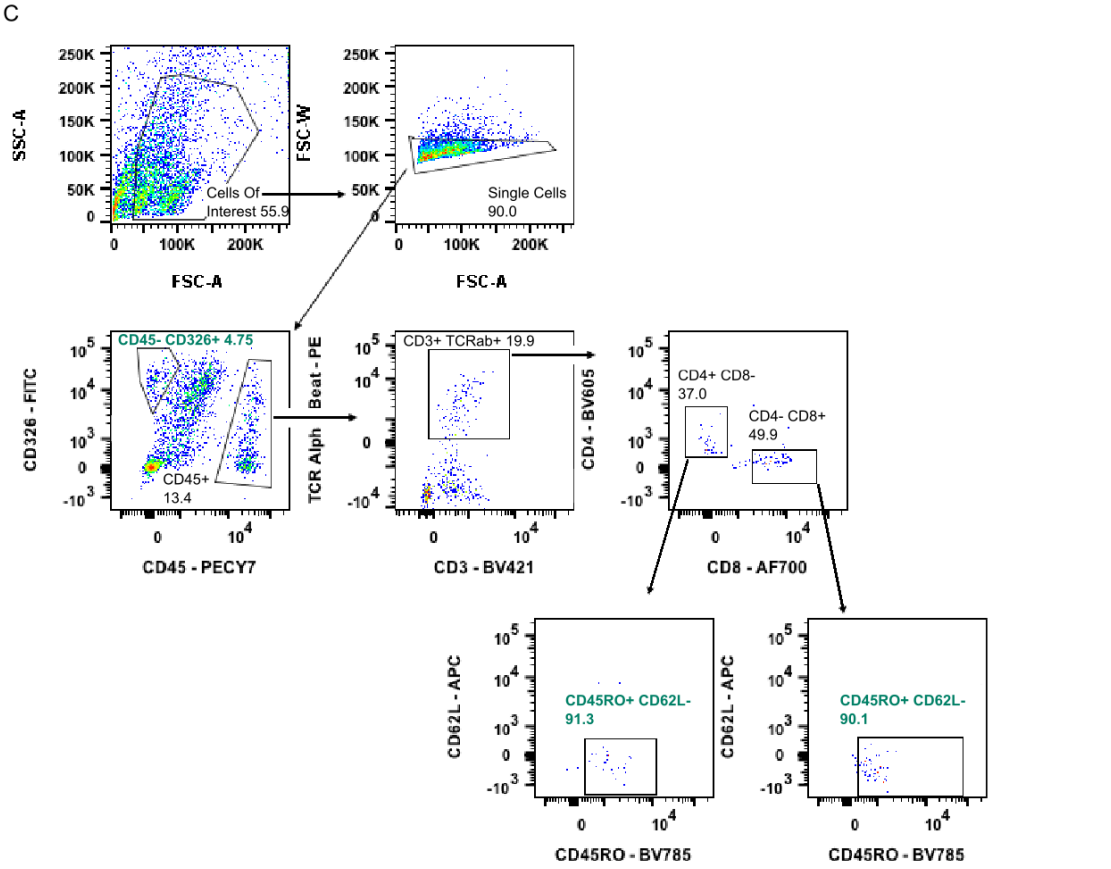


FIGURE CONTINUED IN NEXT PAGE

Figure 3.2 AN EXAMPLE OF FACS GATING STRATEGY USED TO PURIFY INDIVIDUAL CELL POPULATION FROM *A.* BLOOD, *B.* LPL AND *C.* IEL SAMPLES, ALL OBTAINED FROM A SINGLE DONOR. *Populations of interest (for the analysis) are highlight in blue.*

3.5. RNA Sequencing Library Construction And Quality Control

3.5.1 RNA Extraction And Quality Control

RNA from all samples sorted for transcriptomic analysis was extracted by the Qiagen RNeasy micro plus kit (Qiagen, 74034), following the manufacturers instruction. However, instead of mainstream extraction guidelines, the protocol in Appendix D for Total RNA, including small RNA, purification was selected. Extracted RNA quality and quantity was assessed on Agilent Bioanalyzer using the RNA 6000 pico kit (Agilent, 5067-1513).

3.5.2 Smarter Stranded Total RNA Seq Library Construction, Quantification And Quality Control

SMARTer Stranded Total RNA seq Kit- Pico Input Mammalian (Clontech, 635006) was used for all RNA Seq library generation. RNA input was standardized to 800pg, but there were a dozen samples that had less than 800pg of total RNA. In this case, libraries were generated from all RNA available. With exception to fragmentation time, libraries were constructed following the exact manufacturer's instructions. We decreased fragmentation time for samples with RNA integrity number >7 to 3 minutes, as 4 minutes (specified in manual) resulted in over-fragmented inserts.

Sequencing library quality, in terms of insert size distribution and adapter contamination was assessed on Agilent Bioanalyzer via High Sensitivity DNA kit (Agilent, 5067-4626). In addition, mean fragment size for each sample was recorded.

For quantification KAPA Universal Library Quantification kit (KAPABIOSYSTEMS, KK4824) was used following the manufacturer's guidelines. All libraries were diluted 10,000 and 20,000 fold to fall into dynamic range of assay. To ensure the highest accuracy each dilution was run in triplicate with H_20 included as negative control.

QuantStudio 12K Flex Real-Time PCR System (ThermoFisher Scientific, Massachusetts) was used for fluorescence signal (Cq) and melt curve detection. Finally, sample Cq values and associated fragment size were imputed in KAPA Library Quantification Data Analysis Template and concentration of each library calculated.

3.5.3 Library Multiplexing, Pooling and Quality Control

4 to 6 RNA Seq libraries were carefully matched, so that both - UC and Control samples - are represented in one pool. Unfortunately, T cells and B cells were only partially randomized between pools with some residual potential for batch effects.

In addition, extra care was taken to design primer usage and reaction conditions such that all nucleotides at each position of barcodes gets equal representation in each sequencing cycle.

All libraries were normalized to 10nM concentration in Illumina RSB buffer (Kindly provided by sequencing facility) and 10µl of each sample transferred to new tube. Pool quality was assessed on Agilent Bioanalyzer by High Sensitivity DNA kit (Agilent, 5067-4626).

3.6. ATAC Sequencing Library Construction And Quality Control

ATAC Seq assay was modified from the S John *et al* Current Protocols in Molecular Biology 2013 (John *et al.*, 2013) and Buenrostro *et al* Current Protocols in Molecular Biology 2013 (Buenrostro *et al.*, 2013). Due to these modification steps taken, we provide a full procedure description below. Reagents used and buffer recipes are summarized in Table 3.4 and Table 3.5, respectively. The adapter sequences are shown in Table 3.6.

Table 3.4 LIST OF ALL REAGENTS REQUIRED FOR ATAC SEQ LIBRARY CONSTRUCTION.

Reagent	Supplier	Catalogue Number
Trizma® base	Sigma-Aldrich	T6066-1KG
Sodium Chloride	Fisher Scientific	S/3160/53
Potassium Chloride	Sigma-Aldrich	31248
Ethylenediaminetetraacetic acid disodium salt dihydrate	Sigma-Aldrich	E5134 – 1KG
Ethylene glycol-bis(β-aminoethyl ether)-N,N,N′,N′-tetraacetic acid tetrasodium salt	Sigma-Aldrich	E8145-10G
Purified water - Milli-Q® water	Millipore	-
IGEPAL® CA-630	Fluka Analytichal	56741-250ML-F
Illumina Nextra DNA sample prep kit 96 sample	Illumina	FC-121-1031
MinElute Reaction Cleanup Kit	Qiagen	28204
MinElute PCR Purification Kit	Qiagen	28004
DNase/RNase free H20	Qiagen	-
SYBR® Green	Invitrogen	S7585
NEBNext® High-Fidelity 2X PCR Master Mix	New England Biolabs	M0541L
ROX Reference Dye	Invitrogen	12223012
Ethanol Absolute	Fisher Scientific	E/0665DF/17
MB AG 501-X8(D) Resin	BioRad	143-6425
37% Hydrochloric acid	Fisher Scientific	H/1150/PB17
ROCHE- Complete Tablets, EDTA free EASY pack, protease inhibitors	Sigma	04693132001
Spermidine	-	-

Table 3.5 ATAC SEQ BUFFER RECIPES. *Table summarizes all buffers needed for successful library generation. The buffer name is shown in green, where each reagent and its volume are listed below.*

Reagent	Volume	Final Concentration
qPCR Mix		
Nuclease Free Water	4.19µl	
25µM Primer 1 (AD1_noMX_NEXTRA)	0.25µl	
100x SYBR Green	0.06µl	
NEBNext 2x PCR Master Mix	5µl	
PCR Mix		
Nuclease Free Water	8.7µl	
25µM Primer 1 (AD1_noMX_NEXTRA)	2.5µl	
100x SYBER Green	0.3µl	
NEBNext High-Fidelity 2x PCR Master mix	25µl	
ROX Reference Dye	1µl	
Tagmentation Solution (For 50µl Mix) * ***Make Just Before Use and Store on Ice***		
2x Tegmentation DNA Buffer		
Nextra Tagment DNA	2.5µl	
RNase Free Water	22.5µl	
Buffer A:		
Milli-Q Water	47.9ml	
1M Tris.Cl, pH 8.0	750µl	15mM
5M NaCl	150µl	15mM
3M KCl	1ml	60mM
0.5M EDTA, pH 8.0	100µl	1mM
0.5M EGTA, pH 8.0	50µl	0.5mM
0.5M Spermidine	50µl	0.5mM
ROCHE - Complete Tablet (On Day of Experiment)	*1x*	
IGEPAL:		
IGEPAL	4ml (Slightly Warmed)	
Mili-Q Water	36ml (37^{0}C)	
Resin	2g	
** On day of experiment make a serial dilution – 1ml of stock in 9ml Mili-Q Water for 1% stock, then 1:10 in Buffre A for 0.1% stock.*		

Table 3.6 LIST OF ATAC SEQ BARCODE NUCLEOTIDE SEQUENCES (Buenrostro *et al.*, 2013)**.**

Ad1_noMX:	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG
Ad2.1_TAAGGCGA	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGT
Ad2.2_CGTACTAG	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGT
Ad2.3_AGGCAGAA	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT
Ad2.4_TCCTGAGC	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGT
Ad2.5_GGACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGT
Ad2.6_TAGGCATG	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGT
Ad2.7_CTCTCTAC	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGT
Ad2.8_CAGAGAGG	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGT
Ad2.9_GCTACGCT	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGT
Ad2.10_CGAGGCTG	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGT
Ad2.11_AAGAGGCA	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGT
Ad2.12_GTAGAGGA	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGT
Ad2.13_GTCGTGAT	CAAGCAGAAGACGGCATACGAGATATCACGACGTCTCGTGGGCTCGGAGATGT
Ad2.14_ACCACTGT	CAAGCAGAAGACGGCATACGAGATACAGTGGTGTCTCGTGGGCTCGGAGATGT
Ad2.15_TGGATCTG	CAAGCAGAAGACGGCATACGAGATCAGATCCAGTCTCGTGGGCTCGGAGATGT
Ad2.16_CCGTTTGT	CAAGCAGAAGACGGCATACGAGATACAAACGGGTCTCGTGGGCTCGGAGATGT
Ad2.17_TGCTGGGT	CAAGCAGAAGACGGCATACGAGATACCCAGCAGTCTCGTGGGCTCGGAGATGT
Ad2.18_GAGGGGTT	CAAGCAGAAGACGGCATACGAGATAACCCCTCGTCTCGTGGGCTCGGAGATGT
Ad2.19_AGGTTGGG	CAAGCAGAAGACGGCATACGAGATCCCAACCTGTCTCGTGGGCTCGGAGATGT
Ad2.20_GTGTGGTG	CAAGCAGAAGACGGCATACGAGATCACCACACGTCTCGTGGGCTCGGAGATGT
Ad2.21_TGGGTTTC	CAAGCAGAAGACGGCATACGAGATGAAACCCAGTCTCGTGGGCTCGGAGATGT
Ad2.22_TGGTCACA	CAAGCAGAAGACGGCATACGAGATTGTGACCAGTCTCGTGGGCTCGGAGATGT
Ad2.23_TTGACCCT	CAAGCAGAAGACGGCATACGAGATAGGGTCAAGTCTCGTGGGCTCGGAGATGT
Ad2.24_CCACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTGGGTCTCGTGGGCTCGGAGATGT

3.6.1 Nuclear Isolation And Chromatin Tagmentation

First, lysis buffer was made by mixing 0.1% IGEPAL solution 1:1 with Buffer A. Cells were spun at 4°C for 5min at 200rcf and supernatant gently removed from cell pellet. Cells were resuspended in 200µl lysis buffer and left on ice for 8 min exactly to incubate. Following incubation, nuclei were spun at 4°C for 5min at 500rcf. Now, supernatant was gently removed from nuclear pellet, nuclei resuspended in 30µl Tagmentation solution and placed in heating block (at 37°C) for 30 min exactly to incubate. Tagmented DNA was cleaned-up with Quiagen MinElute Reaction Cleanup Kit (elute in 10µl Buffer EB). Tagmented samples were stored at -80°C.

3.6.2 ATAC Seq Library Construction and Quantification

Tagmented DNA samples were removed from freezer and thawed on ice. Next, PCR amplification mix was made by pipetting 37.5 µl of PCR reaction master mix, 2.5µl 25µM Primer 2 (Ad2.1-Ad2.24) and 10µl of transposed DNA in each 1.5 Eppendorf tube. PCR reaction was run as follows:

1 cycle	*5min*	*72°C*
	30 sec	*98°C*
5 cycles	*10 sec*	*98°C*
	30 sec	*63°C*
	60 sec	*72°C*

Hold at 4°C

Sample was removed from PCR machine and keep on ice. qPCR mix was made by combining 9.5 µl of qPCR reaction master mix, 0.5µl 12.5µM Primer 2 (Ad2.1-Ad2.24) and *5µl* mix from 1st PCR in each individual pPCR tube. qPCR reaction was run as follows:

1 cycle	*30 sec*	*98°C*
19 cycles	*10 sec*	*98°C*
	30 sec	*63°C*
	60 sec	*72°C*

Hold at 4°C

Additional cycles for PCR was calculated by plotting Rn vs. cycle. Cycle number that correspond to 25% of maximal fluorescence intensity was used to finish PCR amplification. Libraries were purified using MinElute PCR Purification Kit and stored at -80°C till further analysis.

For ATAC Seq library quantification KAPA Universal Library Quantification kit (KAPABIOSYSTEMS, KK4824) was used. Samples were diluted 1: 10,000 and run in triplicate on QuantStudio 12K Flex Real-Time PCR System (ThermoFisher Scientific, Massachusetts). Concentrations were calculated using constant mean fragment size of 120bp (in line with previous experiments from our lab).

After library submission all concentrations were re-estimated by qPCR-based method, this time using exact fragment size.

3.7 Data Analysis

Initial sequencing data from RNA Seq experiments was processed by the Medimmune Bioinformatics facility, whereas pre-processing of raw ATAC Seq data was performed by Dr. J. Gutierrez-Achurry (postdoctoral research associate in Dr Carl Anderson's group, Wellcome Trust Sanger Institute).

All downstream analysis was performed by the author using either R Studio or Python programming environments. Due unique nature of each analysis pipeline, the detailed analysis steps are outlined in the chapter specific Materials and Methods sections.

3.8 List Of Published Software Packages Used In This Study

Software or Algorithms	Web-access	Published
FASTQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/	Simon Andrews
BWA	http://bio-bwa.sourceforge.net/	(Li and Durbin, 2009)
KEGG	https://bioconductor.org/packages/release/bioc/html/KEGGprofile.html	(Zhao, Guo and Shyr, 2019)
Samtools	http://www.htslib.org/doc/samtools.html	(Li *et al.*, 2009)
Bamtools	https://github.com/pezmaster31/bamtools	(Barnett *et al.*, 2011)
Bedtools	https://bedtools.readthedocs.io/en/latest/	(Quinlan and Hall, 2010)
MACS2	https://github.com/taoliu/MACS/	(Zhang *et al.*, 2008)
Cutadapt	https://cutadapt.readthedocs.io/en/stable/	(Martin, 2011)
PICARD	https://broadinstitute.github.io/picard/command-line-overview.html	-
Sailfish	https://github.com/bcbio/bcbio-nextgen	(Patro, Mount and Kingsford, 2014)
MultiQC	https://multiqc.info/	(Ewels *et al.*, 2016)
Bcbio-nextgen	https://github.com/bcbio/bcbio-nextgen	-
Bcbio-variation	https://github.com/chapmanb/bcbio.variation	-
Novosort	http://www.novocraft.com/products/novosort/	Novocraft Technologies Sdn Bhd
Qualimap	http://qualimap.bioinfo.cipf.es/	(García-Alcalde *et al.*, 2012)
IPA	https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/	-
Cufflinks	http://cole-trapnell-lab.github.io/cufflinks/	(Trapnell *et al.*, 2010)
FeatureCounts		(Liao, Smyth and Shi, 2014)
Sleuth + Wasabi	https://github.com/COMBINE-lab/wasabi/	-
Hisat2	https://ccb.jhu.edu/software/hisat2/index.shtml	(Kim, Langmead and Salzberg, 2015)
Htseq	https://htseq.readthedocs.io/en/release_0.11.1/	(Anders, Pyl and Huber, 2015)
Kraken	https://ccb.jhu.edu/software/kraken/	(Wood and Salzberg, 2014)
R package DiffBind	https://bioconductor.org/packages/release/bioc/html/DiffBind.html	(Stark and Brown, 2011; Ross-Innes *et al.*, 2012)
R package DESeq2	https://bioconductor.org/packages/release/bioc/html/DESeq2.html	(Love, Huber and Anders, 2014)
R package fdrtools	http://www.strimmerlab.org/software/fdrtool/	(Strimmer, 2008)
R package ChipSeeker	https://www.bioconductor.org/packages/release/bioc/html/ChIPseeker.html	(Yu, Wang and He, 2015)
R package TxDb.Hsapiens.UCSC.hg38.knownGene	https://bioconductor.org/packages/release/data/annotation/html/TxDb.Hsapiens.UCSC.hg38.knownGene.html	(Bioconductor Core Team, 2019)
R package org.Hs.eg.db	https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html	(Carlson, 2019)
R package DOSE	https://bioconductor.org/packages/release/bioc/html/DOSE.html	(Yu *et al.*, 2015)
R package ReactomePA	https://bioconductor.org/packages/release/bioc/html/ReactomePA.html	(Yu and He, 2016)
R package clusterProfiler	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html	(Yu *et al.*, 2012)
R package RColorBrewer	https://cran.r-project.org/web/packages/RColorBrewer/index.html	-
R package ggplot2	https://cran.r-project.org/web/packages/ggplot2/index.html	(Ginestet, 2011)
R package rgl	https://cran.r-project.org/web/packages/rgl/index.html	-
R package IRanges	https://bioconductor.org/packages/release/bioc/html/IRanges.html	(Lawrence *et al.*, 2013)

R package factoextra	https://cran.r-project.org/web/packages/factoextra/index.html	-
R package tximport	https://bioconductor.org/packages/release/bioc/html/tximport.html	(Soneson, Love and Robinson, 2016)
R package pheatmap	https://cran.r-project.org/web/packages/pheatmap/index.html	-
R package Hmisc	https://cran.r-project.org/web/packages/Hmisc/index.html	-
R package biomaRt	https://bioconductor.org/packages/release/bioc/html/biomaRt.html	(Durinck *et al.*, 2005, 2009)
R package AnnotationDbi	https://www.bioconductor.org/packages/release/bioc/html/AnnotationDbi.html	(Pagès, Carlson and Falcon, 2019)
R package genefilter	https://bioconductor.org/packages/release/bioc/html/genefilter.html	(Al, 2019)
R package RnaSeqSampleSize	https://bioconductor.org/packages/release/bioc/html/RnaSeqSampleSize.html	(Zhao *et al.,* 2018)
R package topGO	https://bioconductor.org/packages/release/bioc/html/topGO.html	(Alexa A, 2019)
R package plyr	https://cran.r-project.org/web/packages/plyr/index.html	-
R package Rgraphviz	https://www.bioconductor.org/packages/release/bioc/html/Rgraphviz.html	(Hansen *et al,* 2019)
R package Gviz	http://bioconductor.org/packages/release/bioc/html/Gviz.html	(Hahne and Ivanek, 2016)

Resource	**Identifier**
Human Genome GRCh38	http://www.ensembl.org/Homo_sapiens/Info/Index
Human Genome GRCh37	https://genome.ucsc.edu/cgi-bin/hgLiftOver
RStudio	https://www.rstudio.com/
GraphPad Prism 7	https://www.graphpad.com/scientific-software/prism/
Flow Jo	https://www.flowjo.com/
Inkscape	https://inkscape.org/
Python 2 and 3	https://www.python.org/
Microsoft Office	https://products.office.com/
Mac OS Terminal	https://macpaw.com/

4. Comparison And Validation Of Commercially Available Anti-GPR15 Antibodies For Use In Flow Cytometry

4.1 Introduction

Therapies based on the blockade of the intestinal trafficking have shown success in treatment of both - UC and CD (Feagan *et al.*, 2013, Sandborn *et al.*, 2013, Targan *et al.*, 2007, Vermeire *et al.*, 2017). Recently a new G-protein coupled receptor named GPR15 has been reported as an intestine-homing receptor for both human and mice (Fischer *et al.,* 2016, Kim *et al.,* 2013, Nguyen *et al.,* 2015). Despite the efforts, the full functional role of GPR15 has not been yet understood.

Antibodies are immunoglobulin molecules produced by the immune system (Kanyavuz *et al.,* 2019). Antibodies structure allows it to bind to the antigen in a highly specific manner, making them into the one of the most powerful research tools on the market. Antibody production is a very time-consuming process and has a high cost component.

As part of my PhD we asked if any of commercially available anti-GPR15 antibodies could be successfully used in flow cytometry assays, such as immunophenotyping.

4.2 Aim

To compare existing commercially available anti-GPR15 antibodies and validate their performance for immunophenotyping.

4.3 Materials And Methods

4.3.1 Generation Of Jump-In T-REx HEK293 Cell Lines Which Over-Expresses Human GPR15

Jump-In T-REx HEK293 cells were obtained from Medimmune Cell culturing team. Plasmids were designed and given by Matthew Robinson, Medimmune. The full list of reagents and buffers used for transection experiments are summarized in Table 4.1 and Table 4.2, respectively.

Table 4.1 LIST OF ALL REAGENTS REQUIRED FOR JUMP-IN T-REX HEK293 TRANSFECTION EXPERIMENTS.

Reagent	Supplier	Catalogue Number
Geneticin	Gibco	10131-027
Blasticidin	Gibco	A11139-03
Lipofectamine® 2000	Invitrogen	11668-027
Opti-MEM Reduced Serum Medium	Gibco	31985070
DMEM	Gibco	41966-029
GlutaMAX	-	-
NEAA 100X	Gibco	11140-035
FBS dialysed (Tetracycline free)	-	-
R4 integrase	Invitrogen	-
Accutase	Invitrogen	-
PBS	Gibco	-
Vector HuGPR15 16AALVJC in 04-057 K881 1mg/ml	-	-
Vector FlagHuGPR15 16AALVHC in 04-057 K881 1mg/ml	-	-

Table 4.2 SUMMARY OF BUFFER RECIPES NEEDED FOR JUMP-IN T-REX HEK293 TRANSFECTION EXPERIMENTS.

Reagent and Volume	
Transfection Mix (Per Transfection) – Stable Transfection	
Tube A: 150ul OptiMEM +1.25ug of Jump-In plasmid +1.25ug of R4 integrase	
Tube B: 150ul OptiMEM + 6.25 of Lipofectamine 2000	
** When plates (cells) are ready, mix Tube A with Tube B and incubate at room temp for 5 min*	
Transfection Mix (Per Transfection) – Transient Transfection	
Tube A: 250ul OptiMEM +1.25ug of Jump-In plasmid +1.25ug of R4 integrase	
Tube B: 250ul OptiMEM + 6.25 of Lipofectamine 2000	
** When plates (cells) are ready, mix Tube A with Tube B and incubate at room temp for 5 min*	
DMEM Mix	**Final Concentration**
DMEM	
FBS dialysed (Tetracycline Free)	10%
GlutaMAX	1/100
NEAA 100X	1/100

4.3.1.1 Stable Transfections

At Day 1 of transfections, old media was removed and cells incubated with 5ml Accutase at 37°C for 5 minutes. Following incubation, flask was gently taped to dislodge cells and suspension transferred to 50ml Falcon tube, topped up with PBS, spun for 5min at 400g and, finally, resuspend in Opti-MEM for counting. Cells were seeded in 6 well plates at ~1e^{6}cells/well and left overnight to adhere.

At Day 2 - old media was aspirated and replaced with 2ml DMEM Mix without antibiotics (Table 4.2). Cells were transfected using Lipofectamine 2000, where 300ml of transfection mix (Table 4.2) was gently added in each well dropwise. Finally, plates were gently swirled to mix and returned to incubator for next 24h.

At Day 3 - transfected cells were harvested and transferred to new T75 flask containing 25ml of selection median (DMEM Mix + 1mg/ml Geneticin, 5ug/ml Blasticidin). Selection media was changed every 7 days. Once visible colonies had appeared, cells were passaged as normal and used for GPR15 antibody comparison.

4.3.1.2 Transient Transfections

At Day 1 - Jump-in HEK293 were seeded into 6 well plates, so that 625'000cells/well in 2ml DMEM Mix (Table 4.2).

At Day 2 - 500ml of transfection mix (Table 4.2) was gently added in each well dropwise. Finally, plates were gently swirled to mix and returned to incubator for next 24h.

At Day 3 - cells were used for assessment of GPR15 and Flag expression.

4.3.1.3 Plating For Antibody Staining

Old cell media was removed and cells washed with PBS, and incubated with Accutase at 37°C for 3 minutes. Cell suspension was transferred to 15ml or 50ml Falcon tube, topped up with PBS + 1% FCS, spun for 5min at 400g and resuspend in PBS for counting. Cells were seeded in 96 well V-bottom plate at 5.5x10^4 cells/well/100µl.

4.3.2 Peripheral Blood Mononuclear Cell Isolation From Blood Cones.

Ethically sourced blood was purchased from the commercial vendor. Donations were made in day before and blood received in cone-like plastic containers.

Upon arrival, blood cone was cut open and 10ml of blood was run into the 50ml Falcon tube and mixed with the 30ml PBS (Gibco, 10010023). Next, 20ml of the Blood-PBS mix was layered on 20ml of Ficoll (Sigma-Aldrich) and spun for 40min at room temperature at 400g (no brake). PBMC layer was sucked-up via Pasteur pipette, resuspended in 30ml PBS and spun for 10min at 200g. Cell pallet was resuspended in 30ml PBS and spun again for 10min at 200g. As final step, cell pallet was resuspended in 10ml DMEM, counted and seeded at 2e5 cell/well in 96-well U-bottom plate and used for GPR15 antibody assessment.

4.3.3 Antibody Staining

As first step, the eFluor780 viability marker was added into each well (final concentration of 1 in 2500). Plate was wrapped in foil and incubated at either a room temperature (HEK293) or on ice (PBMCS) for 20min. Plate was washed by adding either 100 µl of PBS (HEK293) or cell media (PBMCS) to each well (final 200 µl /well), spun for 5min at 400g and supernatants discarded. Wash step was repeated twice. Figure 4.1. summarizes staining strategy, time and concentrations for all GPR15 experiments. Table 4.3 lists all antibodies used in anti-GPR15 antibody comparison, validation and evaluation experiments.

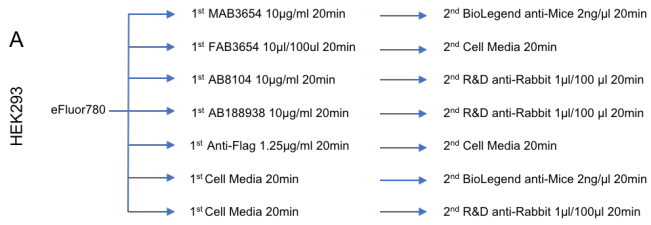


FIGURE CONTINUED IN NEXT PAGE

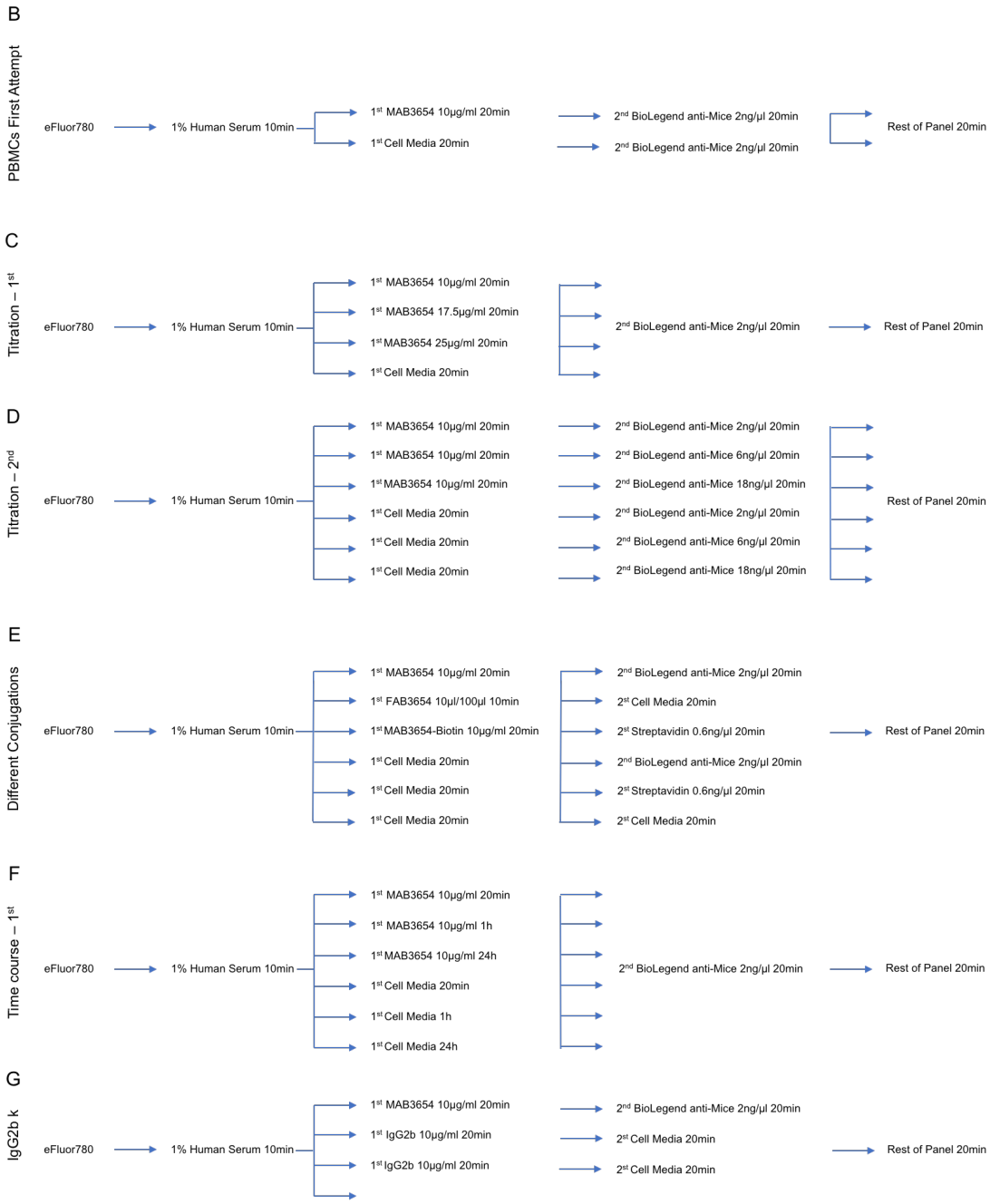


Figure 4.1 DIAGRAMS SHOWING THE STEP BY STEP STAINING STRATEGY FOR EACH INDIVIDUAL ANTI-GPR15 ANTIBODY COMPARISON, VALIDATION AND EVALUATION EXPERIMENT. *Staining strategy for A. commercially available antibody comparison using engineered HEK293 cell lines, B. MAB3654 staining assessment on PBMCs, C. MAB3654 and D. Goat-anti-mice antibody titration, E. tag comparison F. time course and G. isotype control experiments.*

Table 4.3 LIST OF ALL ANTIBODIES USED IN ANTI-GPR15 ANTIBODY EVALUATION EXPERIMENTS.

Marker	Clone	Fluorochrome	Final Concentration	Company	Cat. No.
GPR15	Rabbit Polyclonal	-	10µg/ml	Abcam	AB188938
GPR15	Rabbit Polyclonal	-	10µg/ml	Abcam	AB8104
GPR15	367903	-	10µg/ml	R&D Systems	MAB3654
GPR15	367903	PE	10µl/ml	R&D Systems	FAB3654
GPR15	367903	Biotin	10µl/ml	Conjugated in house	-
Goat-anti-rabbit	Goat Polyclonal	PE	1/100	R&D Systems	F0110
Goat-anti-mice	Goat Polyclonal	PE	2ng/µl	BioLegend	405307
Flag	L5	PE	1.25µg/ml	BioLegend	637310
Viability Dye	-	eFluor 780	1/2500	eBiosciencs	65-0865-14
CD3	UCHT1	PerCP	1/100	BioLegend	300427
CD4	OKT4	BV605	2/100	BioLegend	317437
CD8	HIT8a	AF700	1/100	BioLegend	300919
CD45	HI30	Pacific Blue	1/100	BioLegend	-
Streptavidin	-	PE	0.6ng/ul	eBioscience	-
mIgG2b k	MG2b-57	-	10µg/ml	Biolegend	401201
mIgG2b k	MPC-11	-	10µg/ml	Biolegend	400301

4.3.4 Assessment Of GPR15 Expression At mRNA Level

Blood $CD4^+$ T_H and $CD8^+$ T_C cells were FACS sorted into RLT Buffer (Qiagen, 79216). After cell lysis, RNA was extracted following RNeasy Plus Mini kit (Qiagen, 74134) guidance with extra on column DNase digestion. Next, the High Capacity RNA-to-cDNA Kit (Applied Biosystems™, 4388950) was used to convert RNA into cDNA and gene expression assessed by real-time PCR with TaqMan Gene Expression Master Mix (Applied Biosystems™, 4369016) following the manufacturer's instructions. Both, GPR15 and GAPDH TaqMan probes were kindly given by colleagues in MedImmune. Samples were run and analysed on QuantStudio 12K Flex Real-Time PCR System (ThermoFisher Scientific).

4.3.5 Flow Jo Analysis

For data analysis FlowJo_VX (FlowJo LLC, Ashland) software was employed. Separation Index, calculated as showed below, was used to evaluate the strength of partition between the positive and negative population.

$$Separation\ Index = \frac{MedianPositive - MedianNegative}{(84\%\ Negative - MedianNegative)/0.995}$$

4.4 Results

4.4.1 Comparison Of Commercially Available Anti-GPR15 Antibody Performance Using Genetically Engineered Jump-In HEK293 Cell Lines

We first compared four, at that time commercially available, anti-GPR15 antibodies for their performance on flow cytometry application.

Jump-In HEK293 cell line was engineered to over-express human GPR15 with and without a Flag tag. Transfected cell lines were stained with either MAB3654, FAB3654, AB8104 or AB188938 antibodies followed (where appropriate) by PE-conjugated goat-anti-mouse or goat-anti-rabbit secondary antibodies. Figure 4.2. A shows the initial gating strategy.

AB8104 and AB188938 anti-GPR15 polyclonal antibodies failed to produce any signal when incubated with Jump-In HEK293 cell lines engineered to over-express either human (Figure 4.1. B, $n_{experiment} = 2$, $n_{transfection\ attempts} = 2$) or mice (results not shown) GPR15 with and without Flag tag.

MAB3654 and FAB3654 showed weak signal when incubated with cell lines transfected with GPR15 alone (Figure 4.1.B, $n_{experiment} = 2$, $n_{transfection\ attempts} = 2$), yet no signal was observed when cell lines engineered to overexpress both GPR15 and Flag tag was tested.

However, the Flag tag expression was very low (~10%) (Figure 4.1 B) suggesting that there are underlining issues with transfection itself. Therefore, the strength of signal produced by any of four anti-GPR15 antibodies cannot be used as definitive quality assessment.

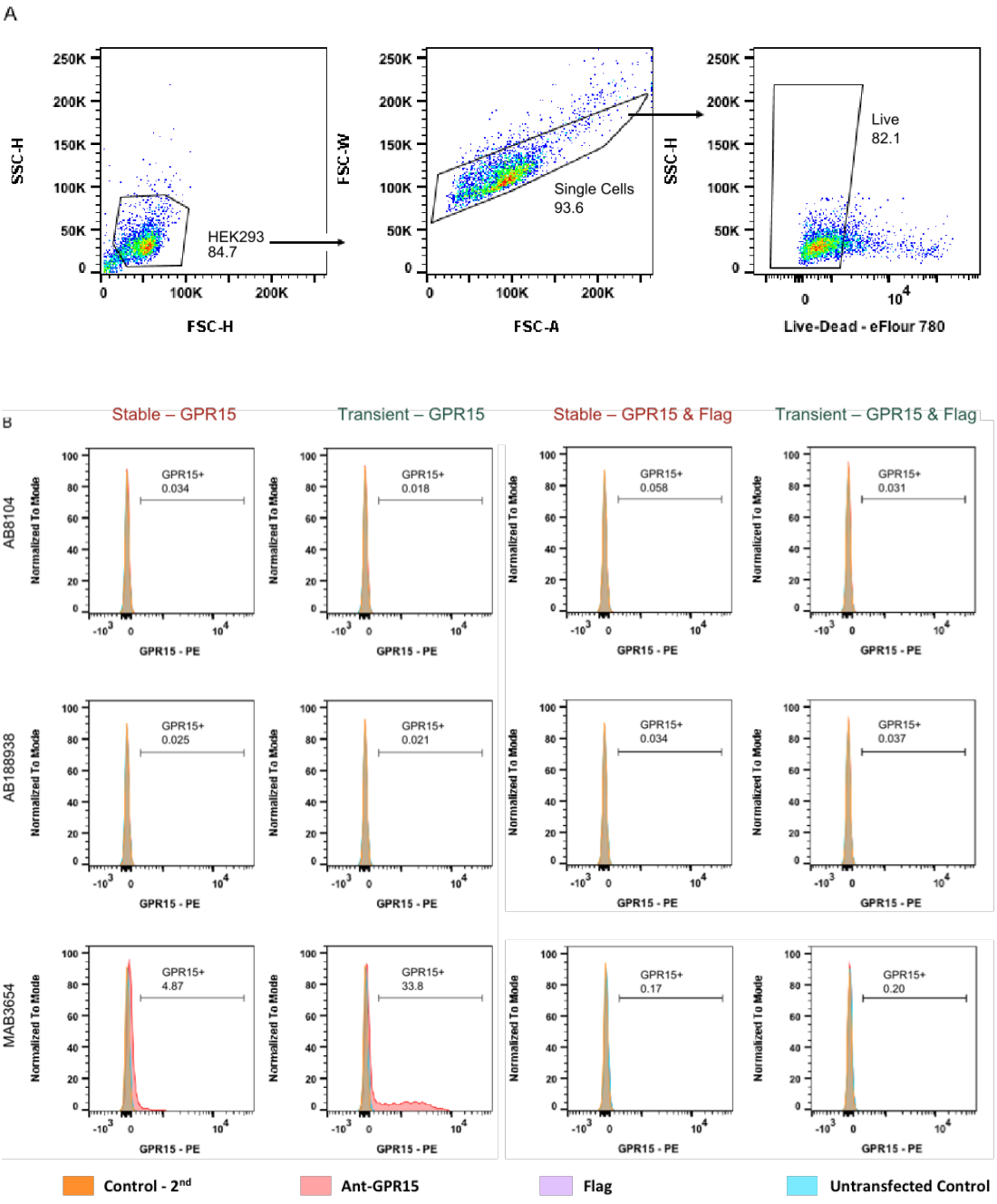


FIGURE CONTINUED IN NEXT PAGE

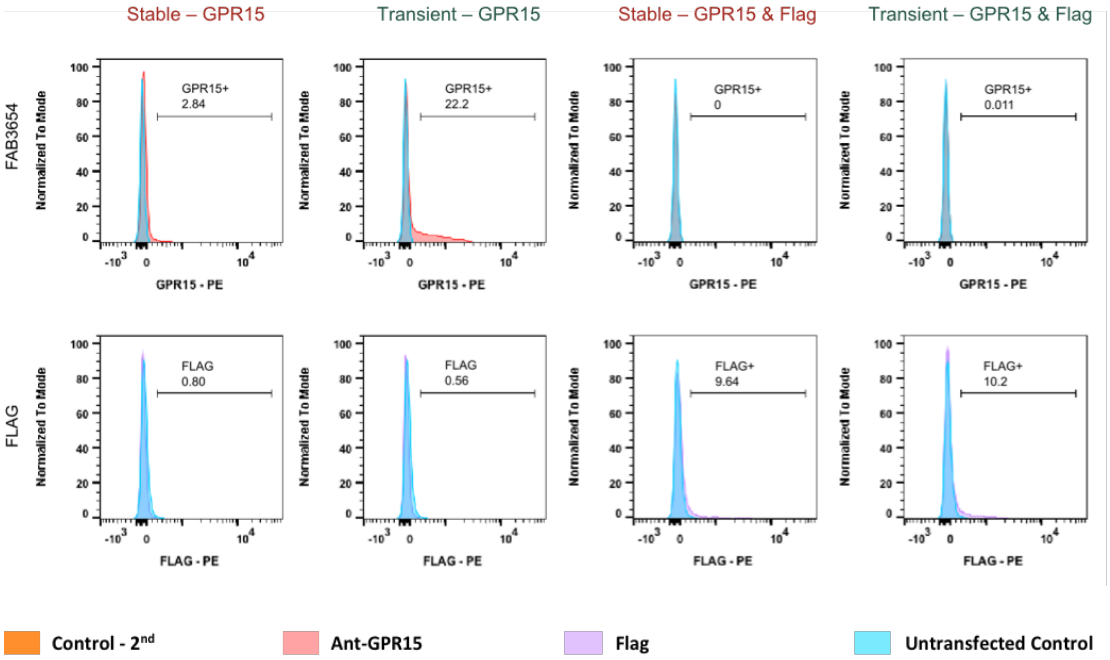


Figure 4.2 COMPARISON OF COMMERCIALLY AVAILABLE ANTI-GPR15 ANTIBODIES *($n_{experiment}$ = 2, $n_{transfection\ attempts}$ = 2). Flow cytometry plots showing the A. initial gating strategy and B. anti-GRP15 antibody (AB8104, AB188938, MAB3654 and FAB3654) and anti-Flag staining performance. Histograms show PE fluorescence normalized to mode.*

4.4.2 Assessment Of MAB3654 Staining On Peripheral Blood T Cells

Antibody performance is subject to both - an application its intended to be used and the cellular context. At this stage, before investing time in an extensive antibody validation, we wanted to see if MAB3654 would show any binding that could be detected by flow cytometry when incubated with immune cells isolated from the blood.

PBMCs were stained with MAB3654 followed by PE-conjugated goat-anti-mice secondary antibody. Gating was set by looking at the signal produced by the secondary antibody alone (2^{nd} alone contains all markers, except the MAB3654). Figure 4.3 A shows the initial gating strategy.

Flow cytometry analysis showed that MAB3654 has an ability to bind both - $CD4^+$ and $CD8^+$ T cells. However, there was a marked inter-donor variation in GPR15 expression by $CD8^+$ T_C (Figure 4.3 B) and $CD4^+$ T_H (results not shown) cells.

In summary, both $CD8^+$ and $CD4^+$ T lymphocytes stained positive for GPR15 when incubated with MAB3654.

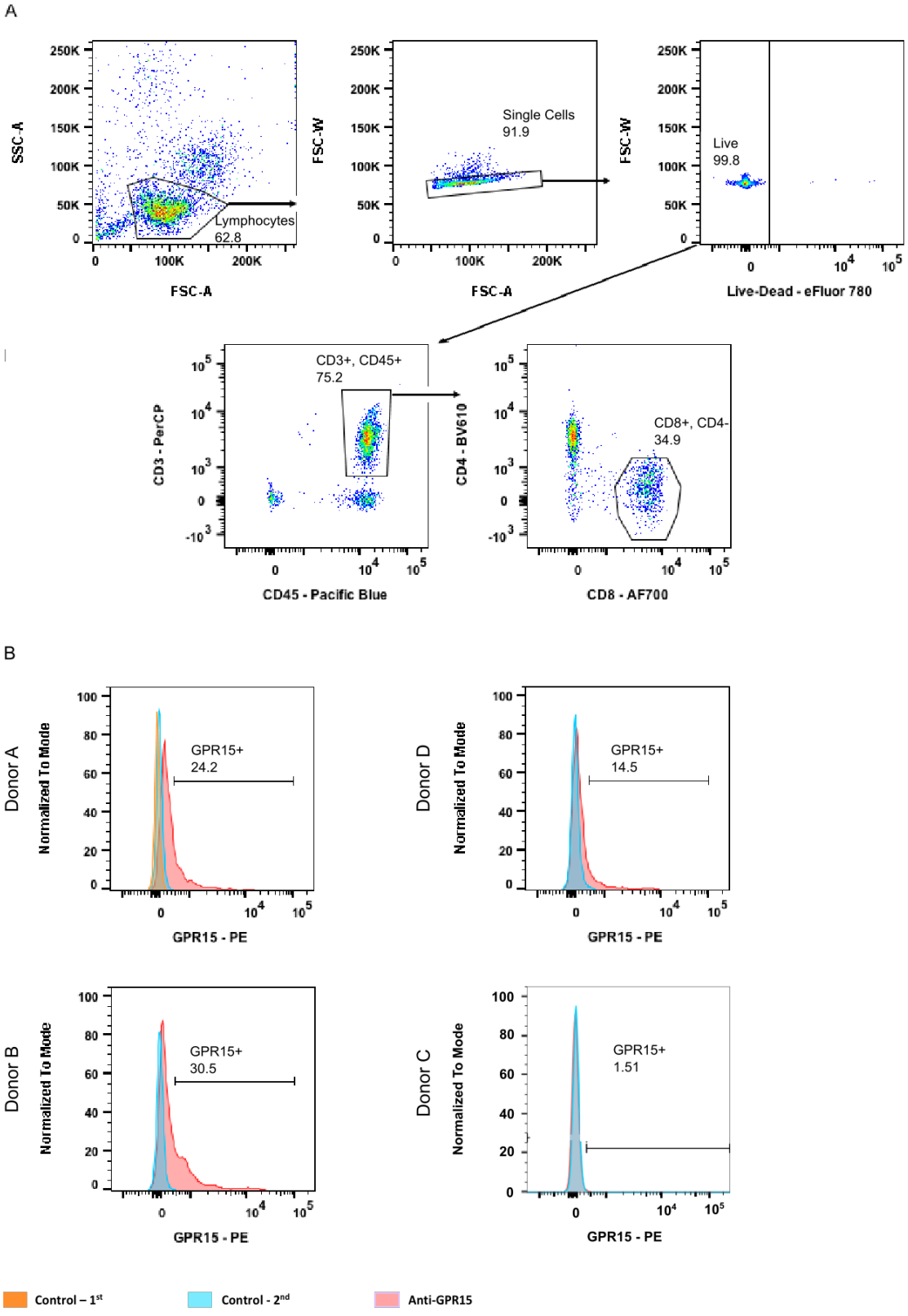


Figure 4.3 FLOW CYTOMETRY PLOTS SHOWING THE ANTI-GRP15 ANTIBODY MAB3654 STAINING PERFORMANCE (*$n_{experiment} = 2$, $n_{donor} = 4$). Flow cytometry plots showing the A. initial gating strategy and B. GPR15 expression by blood CD8 T cells. Histograms show PE fluorescence normalized to the mode.*

4.4.3 MAB3654 Staining Optimization For Flow Cytometry Using Blood Resident T Cells

In our previous experiment we asked if MAB3654 would show any signal when incubated with blood resident T cells. Indeed, we observe some binding, but it displayed large inter-individual variation. Here we set out to assess if by tweaking the experimental design a more reproducible stain could be achieved. Figure 4.4 A shows initial gating strategy used for all optimization experiments.

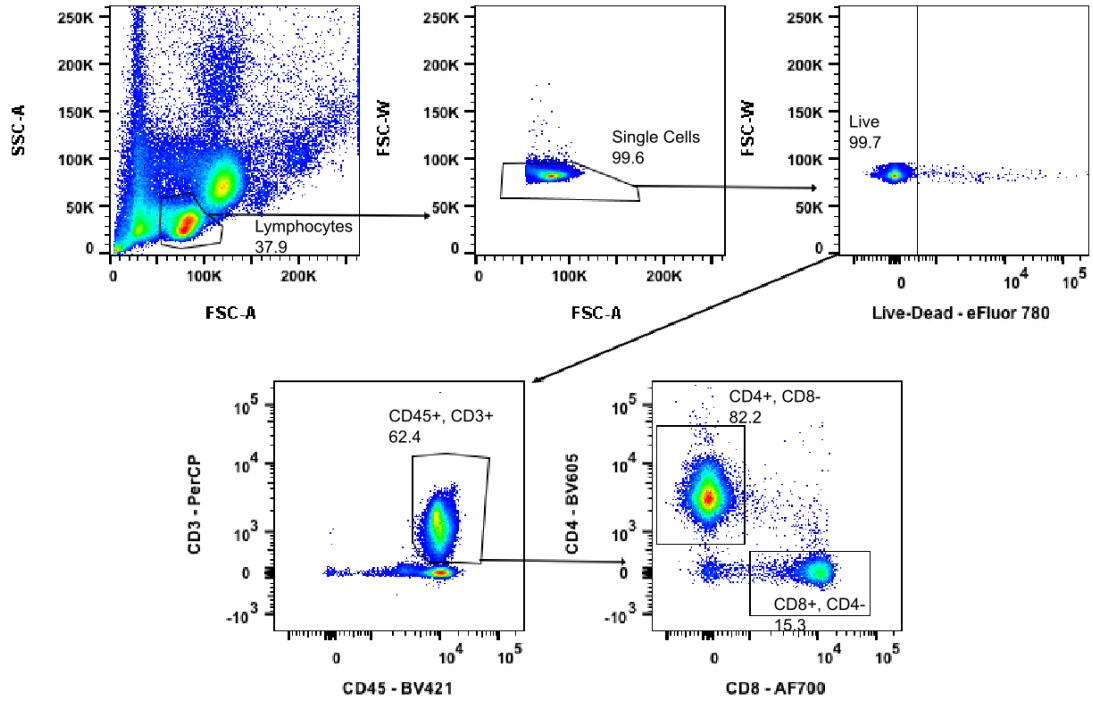
To determine an optimal antibody staining concentrations, we performed serial dilution of both MAB3654 primary and PE-conjugated goat-anti-mice secondary antibodies. Separation between the positive and negative populations was measured by SI.

The best distinction between GPR15 positive and negative population was achieved when MAB3654 (conc. = 17.5mg/ml). However, all 3 concentrations tested produced visually easy-to-separate dot-plots with almost equal % of $GPR15^+$ events ($n_{experiment} = 1$, $n_{donor} = 1$, Figure 4.4 B). PE-conjugated goat-anti-mice secondary antibody showed the best performance at conc. = 0.2ug/100ul. Higher concentrations (≥0.6ug/100ul) of goat-anti-mice secondary antibody, in the absence of MAB3654, resulted in an unspecific binding ($n_{experiment} = 1$, $n_{donor} = 1$, Figure 4.4 C). The amount of unspecific binding of goat-anti-mice secondary antibody, when MAB3654 was present, was very hard to estimate as all 3 concentrations had similar % of $GPR15^+$ events.

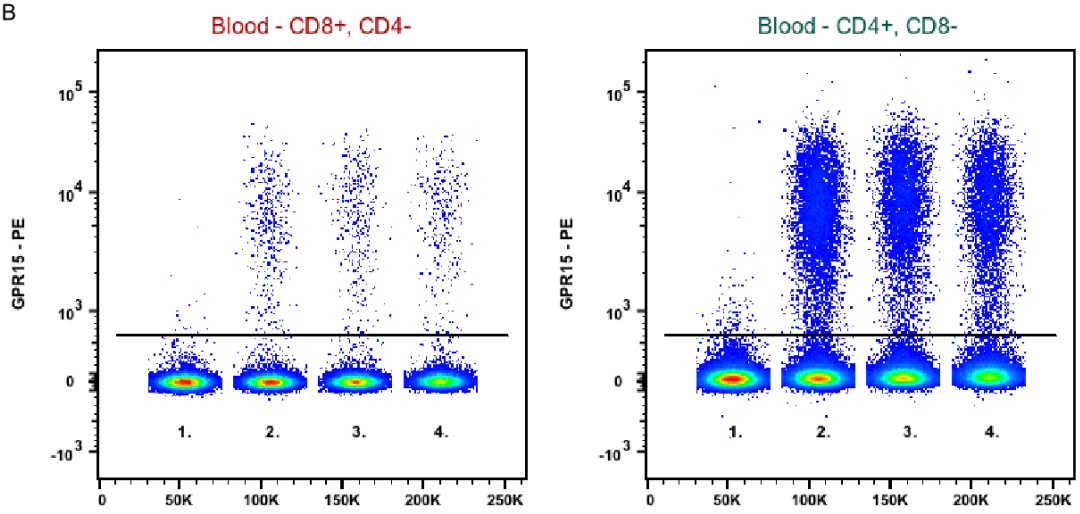
To further evaluate if goat-anti-mice secondary antibody shows some unspecific binding when in prescience of MAB3654, we performed side by side comparison of staining quality when the same anti-GPR15 antibody clone was used already conjugated with a PE (FAB3654) or biotin (produced in MedImmune). Very similar GPR15 expression levels were detected by all three methods ($n_{experiment} = 1$, $n_{donor} = 3$, Figure 4.4 D), but the "indirect" staining method with MAB3654 followed by PE-conjugated anti-mice secondary showed by far the best separation (SI_{Tc} = 97.9, SI_{TH}= 76.9).

In addition to determining the optimal antibody staining concentration, we assessed if longer incubation times would lead to better distinction between the positive and negative populations. PBMCs were stained with MAB3654 for 20min, 1h or 24h (at dark on ice). Even though MAB3654 incubation time of 1h showed the best SI (*SI = 92*), both 20min and 1h stains were visually easy to separate and returned almost identical % of GPR15^{+} events ($n_{experiment} = 1$, $n_{donor} = 1$, Figure 4.4 E).

A



B

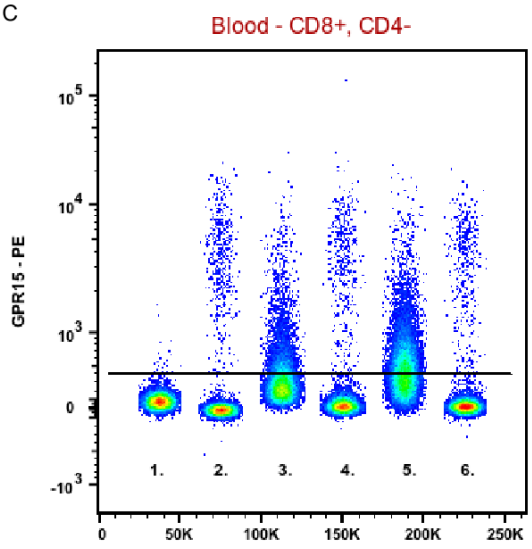


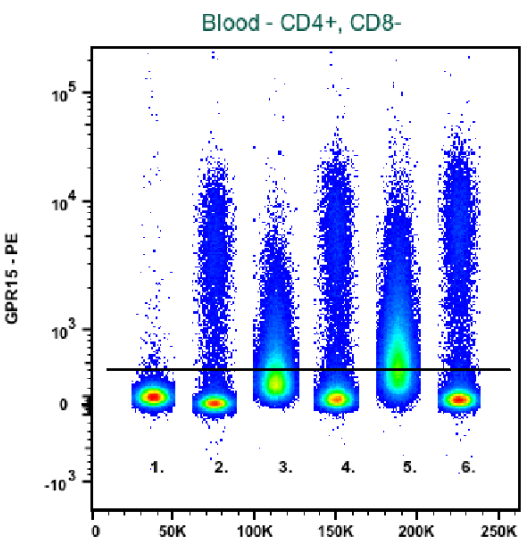
Sample – CD8	1.	2.	3.	4.
Stain	2nd	Full	Full	Full
1st Conc.	10mg/ml	10mg/ml	17.5mg/ml	25mg/ml
GPR15+	0.06	1.58	1.52	1.63
SI		115	132	126

Sample - CD4	1.	2.	3.	4.
Stain	2nd	Full	Full	Full
1st Conc.	10mg/ml	10mg/ml	17.5mg/ml	25mg/ml
GPR15+	0.11	6.82	6.56	6.84
SI		112	123	104

FIGURE CONTINUED IN NEXT PAGE

C

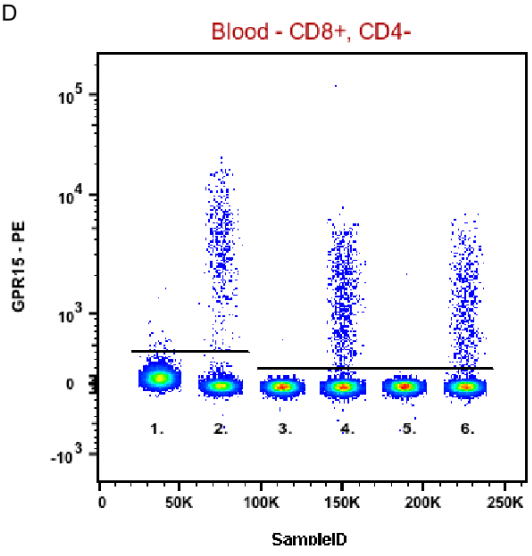


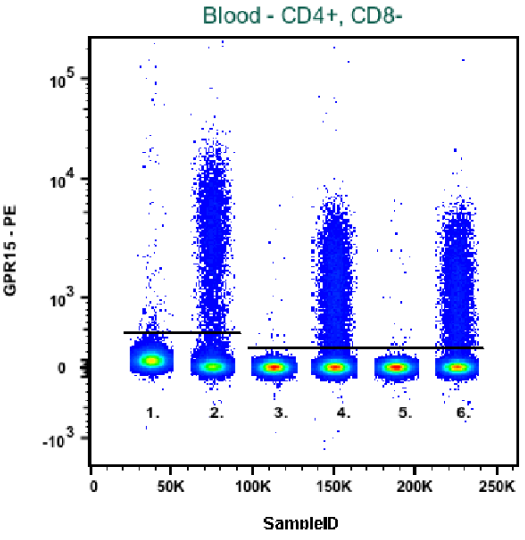


Sample – CD8	1.	2.	3.	4.	5.	6.
Stain	2nd	Full	2nd	Full	2nd	Full
2ND Conc.	0.2µ in 100µl	0.2µ in 100µl	0.6µ in 100µl	0.6µ in 100µl	1.8µ in 100µl	1.8µ in 100µl
GPR15+	0.22%	4.67%	23.3%	3.97%	50.3%	3.10%
SI		97.9		65.6		82.4

Sample – CD4	1.	2.	3.	4.	5.	6.
Stain	2nd	Full	2nd	Full	2nd	Full
2ND Conc.	0.2µg in 100µl	0.2µg in 100µl	0.6µg in 100µl	0.6µg in 100µl	1.8µg in 100µl	1.8µg in 100µl
GPR15+	0.27%	8.55%	30.4%	9.6%	66.6%	8.8%
SI		76.9		54.4		87.8

D





Sample – CD8	1.	2.	3.	4.	5.	6.
Stain	2nd	Full	2nd	Full	2nd	Full
Tag	Mab + 2nd	Mab + 2nd	Biotin	Biotin	Fab	Fab
GPR15+	0.22	4.67	6.84E-3	4.83	6.63E-3	4.16
SI		97.9		33.8		32.7

Sample – CD4	1.	2.	3.	4.	5.	6.
Stain	2nd	Full	2nd	Full	2nd	Full
Tag	Mab + 2nd	Mab + 2nd	Biotin	Biotin	Fab	Fab
GPR15+	0.27	8.55	0.033	7.81	0.04	7.71
SI		76.9		28.7		29.0

FIGURE CONTINUED IN NEXT PAGE

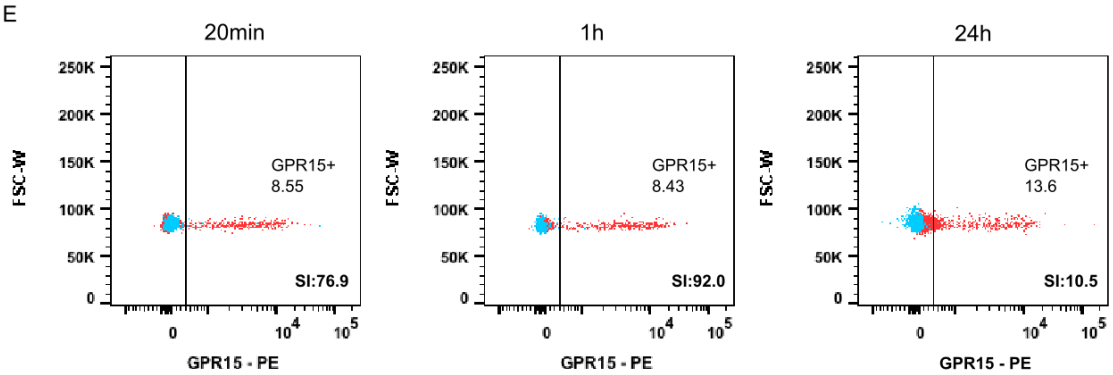


Figure 4.4 FLOW CYTOMETRY PLOTS SHOWING THE *A.* INITIAL GATING STRATEGY, *B.* MAB3654 PRIMARY AND *C.* GOAT-ANTI-MICE SECONDARY ANTIBODY TITRATION, *D.* TAG COMPARISON AND *E.* MAB3654 INCUBATION TIME COMPARISON. *Histograms show PE fluorescence normalized to the mode.*

4.4.4 MAB3654 Staining Validation Using Blood Resident T Cells

Probably the most fundamental characteristics of an antibody is its ability to bind selectively to its antigen. It was crucial for us to determine if GPR15 positive population is truly expressing GPR15. We used flow cytometry to sort GPR15^{-}, GPR15^{LOW} and GPR15^{HIGH} T cells. *GPR15* expression (at mRNA level) was assessed by TaqMan gene expression assay, normalized against the GAPDH and ratio calculated by $2^{-\Delta\Delta Ct}$.

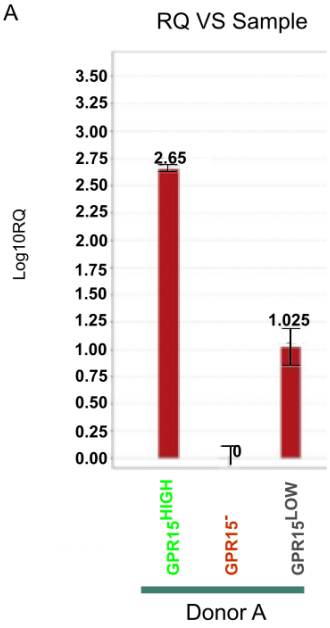
The GPR15^{LOW} and GPR15^{HIGH} populations had 10.591 and 446.389 times higher *GPR15* expression than the GPR15^{-} negative population, showing that MAB3654 truly binds to the GPR15 ($n_{experiment} = 1$, $n_{donor} = 4$, Figure 4.5 A).

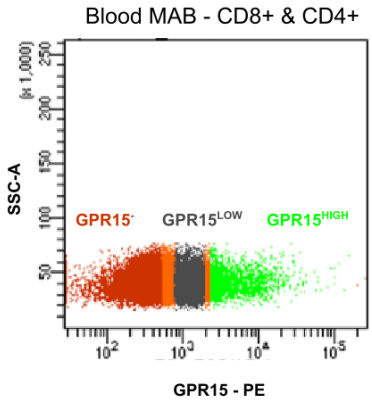
Next, we used an isotype control to assess the level of non-specific binding. PBMCS were stained with either MAB3654 or IgG 2b isotype controls (two different clones were given by colleagues in MedImmune), followed by goat-anti-mice secondary antibody. Both isotype controls showed very little signal ($n_{experiment} = 1$, $n_{donor} = 1$, Figure 4.5 B).

Unfortunately, due unforeseeable reasons project was stopped at this stage and no other validation experiments carried out.

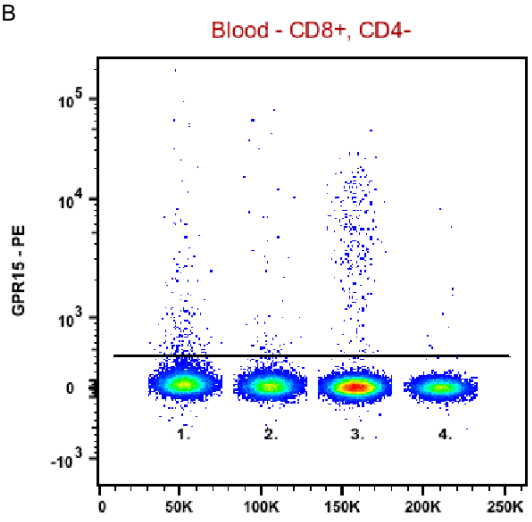
In summary, we showed that cells bound by the MAB3654 were enriched for *GPR15* expression.

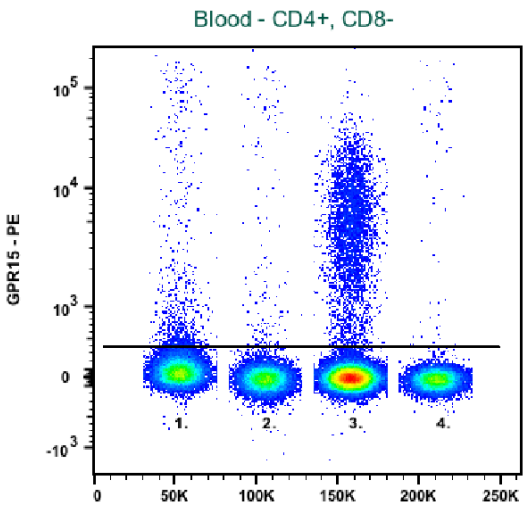
A





B





Sample – CD8	1.	2.	3.	4.
Stain	IgG 2b (A)	IgG 2b (B)	Full	2nd
GPR15⁺	1.48	0.51	0.88	0.11

Sample – CD4	1.	2.	3.	4.
Stain	IgG 2b (A)	IgG 2b (B)	Full	2nd
GPR15⁺	3.81	0.95	6.38	0.38

Figure 4.5 MAB3654 STAINING VALIDATION WITH *A.* qPCR ($n_{experiment} = 1$, $n_{donor} = 4$) **AND *B.* IgG 2b ISOTYPE CONTROL** ($n_{experiment} = 1$, $n_{donor} = 1$)**.**

4.5 Discussion

In early experiments from our laboratory (carried out by Josquin Nyce) we used GPR15 positive (HDLM-2) and negative (MOLT4) cell lines, selected based on information in The Human Protein Atlas. Only very faint shift in GPR15 expression was detected and it could not be distinguished from an artefact. We reasoned that HDLM-2 expresses too little GPR15 at protein level and, therefore, a cell line engineered to over-express GPR15 would be more informative for further evaluation of antibody performance.

In the first part of this study, we transfected Jump-In HEK293 cell lines to over-express GPR15 alone or with Flag tag. We performed both stable and transient transfections with aim to ensure that we achieve noticeable over-expression of our target protein. Only MAB3654 and FAB3654 produced any signal when incubated with cell lines transfected to over-express GPR15 alone. As expected the signal from transiently transfected cell line was much higher than from stable transfections, yet still below to the expected transfection efficiency. Indeed, the very low expression of Flag tag (~10%) by cell lines engineered to over-express the GPR15 with the Flag tag suggest that there is an underlying issue with the transfection itself. Therefore, the signal levels produced by any of the four antibodies could possibly be attributed to an experimental downfall and not the antibody. We currently do not know why transfections did not work and it was not in scope of our immediate goals to investigate it further.

Our main goal was to have an anti-GPR15 antibody that could be used for immunophenotyping applications. From antibodies tested, only MAB3654 had shown some degree of signal in both HDLM-2 and transfected HEK293 cell lines. However, before proceeding to time-consuming MAB3654 validation experiments, we wanted to see if MAB3654 would produce any signal when incubated with blood resident T lymphocytes. The initial experiments showed high inter-individual variation in GRP15 expression by both $CD8^+$ T_c and $CD4^+$ T_H cell. Closer assessments of Flow Jo dot plots showed that for some of the donors' "negative" population has become elongated, which could be indicative for either - a cell subpopulation that has very low GPR15 expression or an issue with the staining protocol itself.

Therefore, we proceeded to investigate if further optimization of our staining protocol could resolve variation in MAB3654 staining. We observed that higher concentration and longer incubation time of MAB3654 resulted in better SI. However, visual assessment of flow cytometry graphs showed no noticeable difference in staining quality between the optimized experimental settings and our initial staining protocol. We would like to acknowledge that our antibody titration experiments were not comprehensive. Our goal was to see if substantial increase in antibody concentration would increase the target detection. Therefore, if MAB3654 passed the performance assessment, more extensive titration should be carried out.

Interestingly, all further experiments including the optimization experiments showed very consistent anti-GPR15 staining ($\%\ GPR15^{+}{}_{Tc} = 2.58 \pm 0.85$, $\%\ GPR15^{+}{}_{TH} = 6.5 \pm 1.83$; $n_{experiment} = 5$, $n_{donor} = 8$). The only difference we could identify between our initial PBMC staining and optimization experiments was new batch of the goat-anti-mice secondary antibody.

Our results are supported by recent publication by *Adamczyk et al., 2017*. They used FAB3654 to immunophenotype T lymphocytes from the peripheral blood of healthy donors and UC patients. Similar to us, they showed that only small percentage of healthy $CD4^{+}$ and $CD8^{+}$ T cells expresses the GPR15 ($\%\ GPR15^{+}{}_{Tc} = 1.9 \pm 1.0$, $\%\ GPR15^{+}{}_{TH} = 4.0 \pm 2.7$). However, it is important to highlight a paper form Bauer *et al* where they showed that history of tobacco-smoking can have a severe effect on GPR15 abundance. In their study, current smokers had a markedly higher proportion of GPR15 positive T cells (10%-30%) than the healthy non-smokers (median 5%) (Bauer *et al*., 2017). The smoking history of blood donors was not provided. However, it would be interesting to determine if the high inter-individual variation in GRP15 expression observed in initial experiments was due the tobacco-smoking.

In the light of ongoing concerns about data reproducibility, it is of high importance to ensure that reagents, such as antibodies are subjected to rigorous application validation, before use in research. For antibody to be used in research it should be specific,

selective and reproducible. As a final part of this study, we sought to get an experimental proof that MAB3654 is suitable for immunophenotyping application.

We first performed a qPCR on FACS sorted T cell (binned by different levels of GPR15 expression) and showed that MAB3654 truly separates GPR15 positive events (at mRNA level). Next, we used an isotype control to assessed the level of non-specific staining, such as, antibody binding to the FC receptors. We showed that there is very little non-specific binding. However, as only a small percentage of blood resident T_C and T_H cells expressed GPR15, even very little of non-specific binding could lead to large proportion of false positives. We have been pre-incubating our cells with Human serum for 10min before staining, but it should be tested if increase in blocking time could reduce the non-specific binding to even lower levels.

Unfortunately, project needed to be stopped and we could not do any further validation or protocol optimization experiments. However, if more time would be given, we would like to test MAB3654 specificity for GPR15. The one of experiment that could be done to evaluate MAB3654 specificity, is to transfect a cell line to express closely related family members of GPR15. However, the antibodies selectivity for GPR15 in over-expressed system might be different from one in physiological expression densities. Hereby a different approach to assess the MAB3654 specificity and sensitivity would be to use gene editing tools to knock-down the expression of GPR15.

In summary, we evaluated performance of four commercially available anti-GPR15 antibodies and rough assessment showed that MAB3654 has the most potential to be used as research tool for our immunophenotyping experiments.

5. Assessment Of Differences In Transcriptional Profile Between The Healthy Controls And UC Patients

5.1 Introduction

To date most attempts to describe the transcriptional landscape in UC have relied upon data from whole tissue samples (Costello *et al.*, 2005; Planell *et al.*, 2013; Bjerrum *et al.*, 2014; Van der Goten *et al.*, 2014; Taman *et al.*, 2018; Haberman *et al.*, 2019), where cell heterogeneity and dynamic ratios hinders identification of transcriptional changes associated with disease and subsequent gene – risk associated variant interplay. In this context, it is important to consider studies such as those by *Fairfax et al* and *Dimas et al* which have highlighted that the impact of genetic variants on gene expression may be cell-type dependent and also depend upon inflammatory state (Dimas *et al.*, 2009; Fairfax *et al.*, 2012, 2014). Nevertheless, for most of immune-mediated diseases, including UC, the pathological cell types are unknown. The observations that IBD associated causal variants impair genes acting in various inflammatory pathways (Pidasheva *et al.*, 2011; Rivas *et al.*, 2011a; Bouzid *et al.*, 2013; Zhu *et al.*, 2017) and disease pathology is driven by an uncontrolled immune response, support a key role for immune cells.

In this chapter, we employed RNA Seq to investigate transcriptional changes associated with different disease states or anatomical locations in a cell type specific manner.

5.2 Aims

- To compare the transcriptional activity in purified immune cell populations from healthy volunteers and UC patients with or without active inflammation in their sigmoid colon.

- To assess how does the transcription profile changes in the same cell type depending on its anatomical location.

5.3 Materials And Methods

5.3.1 Sequencing Design

Before sequencing all samples collected during the study, we decided to perform optimization experiments to assess the sequencing library quality and identify the optimal sequencing design (For more detail see Appendix 1). We found that libraries made from samples low in RNA quantity showed poor alignment to the human genome and introduced RNA quality and quantity-based library pre-filtering (for more detail see Appendix 2).

92 libraries were sent for sequenced to SMCL Next Generation Sequencing Hub, Addenbrookes, Cambridge. Samples were sequenced on Illumina HiSeq 4000 at 110M 75bp PE reads/library. 15% Phix spike in was added for each run.

5.3.2 RNA Seq Data Analysis Pipeline

Initial quality assessment and read mapping was performed by the MedImmune Bioinformatics facility, whereas further downstream analysis was carried out by the author.

5.3.2.1 Pre-Processing Of Raw Sequencing Data

The initial analysis was performed using the *bcbio* open resource python toolkit. The initial analysis step by step included:

- Raw sequencing read quality control by *FastQC* quality metrics tool;
- Removal of left-over adapters;
- Raw read alignment to the human reference genome (hg38) by *Hisat2* (Kim, Langmead and Salzberg, 2015) and pseudoalignment to the human reference transcriptome by *Sailfish* (Patro, Mount and Kingsford, 2014)*;*
- Alignment quality control by *Samtools* (Li *et al.*, 2009) and *Qualimap* (García-Alcalde *et al.*, 2012)*;*

The end product of this initial pipeline was Sailfish files, containing transcript level alignment for each RNA library sequenced along with a *MultiQC* report summarizing results from all individual QC steps included in workflow.

5.3.2.2 Downstream Analysis

5.3.2.2.1 Creation Of Gene-Level Count Datasets

The *tximport* package (Soneson, Love and Robinson, 2016) was used to summarize *Sailfish* generated transcript level abundance into gene-level expression count matrices. Next, newly generated gene-level count matrices and associated metadata were passed to the *DESeq2* R package (Love, Huber and Anders, 2014) selected for differential expression calculation. Finally, genes with no counts or single count across all samples were removed.

5.3.2.2.2 Gene And Sample Sub-Setting

Protein coding and lncRNA gene lists were obtained from Ensemble data base using *biomaRt* R package (Durinck *et al.*, 2005). Then, both lists, one by one, were used to filter our data set and create two new data matrices. One with only protein coding genes and second with both lncRNA and protein coding genes.

Data QC showed that all cell populations extracted from LPL had much higher intra-group variance than their blood relatives (Appendix 3). Disproportionate variance could introduce a bias in the test statistics if the differential expression call would have been made on the whole dataset. Therefore, both protein coding and protein coding + lncRNA data sets were split into smaller data matrices. Samples in each newly created object are listed below.

- LPL $CD4^+$ T_{EM} from UC_i and UC_n and C
- LPL $CD19^+$ B cells from UC_i and UC_n and C
- Blood $CD4^+$ T_{EM} from UC_i and UC_n and C
- Blood $CD19^+$ B cells from UC_i and UC_n and C
- Blood $CD4^+$ T_{EM} from C and Blood $CD19^+$ B cells from C
- Blood $CD4^+$ T_{EM} from C and LPL $CD4^+$ B cells from C
- Blood $CD19^+$ T_{EM} from C and LPL $CD19^+$ B cells from C

5.3.2.2.3 Sample Quality

We put large emphasis on data quality assessment. An extensive pre- and post- differential expression QC was performed, including the Power calculation to assess if our data set is sufficient in size to confidently reject the null hypothesis. For in-depth description of the QC, Power calculation and assessment of main challenges encountered please refer to Appendix 1-5.

5.3.2.2.4 Call For Differential Expression

During QC we noticed that many of smaller data matrices were below the Independent filtering threshold. Meaning that genes with very low counts were retained. Therefore, we introduced 2 different gene-count pre-filtering functions for protein coding and protein coding + lncRNA data sets, respectively.

Protein Coding Data Set

Before calling for differential expression, all sample subsets were filtered to remove genes where there were less than *n* number of samples with normalized counts greater than or equal to 10. n was unique for each subpopulation and defined by taking a half of sample number available within the phenotype with lowest sample numbers.

For differential expression analysis of protein coding genes, a *design* formula was set on either:

- RNA sample collection time and disease state (e.g. C vs UC_i, C *vs* UC_n)
- RNA sample collection time and anatomical location (e.g. blood *vs* LPL)
- RNA sample collection time and cell type (e.g. CD4 *vs* CD19)

Protein Coding + lncRNA Data Set

The gene-count pre-filtering threshold was increased so that genes would be retained if all samples from a given subset expressed more than 10 normalized counts each. Sex was added to *design*.

Protein Coding And Protein Coding + LncRNA Data Set

After calling for differential expression, test statistics in terms of p-value distribution, outlier removal, independent filtering and differentially expressed gene count distribution was checked. *fdrtool* R package (Strimmer, 2008) was used to determine if populations compared showed different null variance than expected by statistical tests implemented in *DESeq2*. For full details please see the Appendix 3.

AnnotationDbi and *org.Hs.eg.db* R packages (Carlson, 2019; Pagès, Carlson and Falcon, 2019) were used to assign the gene symbols and Entrez ID to all genes used for differential expression calculation.

5.3.2.2.5 Pathway Analysis

We performed Ingenuity Pathway Analysis (IPA) (Qiagen) and Gene Ontology (GO) enrichment analysis of differentially expressed genes. For IPA analysis background was set to all genes expressed by population of interest, obtained from DESeq2 expression matrix. GO enrichment was determined by *topGO* R package (Alexa A, 2019). The GO background was set so that only comparison specific genes which matched in expression strength with DEG was used. In both cases enrichment was determine by Fisher's statistical test. For GO analysis the node size was set to 5.

5.3.2.2.6 Overlay With Published Literature

We compared the DEG list identified in our study to already published studies by *Taman et al* and *Van der Goten et al* (Van der Goten *et al.*, 2014; Taman *et al.*, 2018). First, lists of differentially expressed genes was downloaded from supplementary information attached to the publication. Next, we used *biomaRT* to convert the gene symbols (Taman *et al.*, 2018) and Affymetrix GeneChip Human Gene 1.0St probe IDs (Van der Goten *et al.*, 2014) to human Hg38 Ensemble IDs. Finally, we used Ensemble IDs to filter out the matching genes pairs.

Manual PubMed search for each DEG (with median normalized count >100) was carried out. Search criteria was set to either gene symbol + Ulcerative colitis, gene symbol + Inflammatory Bowel Disease or gene symbol + cell type (e.g. T cells/B cells). A hit is

term given for a DEG that returned at least one publication. However, the literature search is a based-on term chain appearing in text independent on the context of publication.

5.4 Results

5.4.1 Determining Disease Specific Change In Expression Profiles In Purified CD19$^+$ B Cells And CD4$^+$ T_{EM} Immune Cell Populations From Peripheral Blood And Sigmoid Colon

We hypothesised that UC associated risk variants indirectly lead to the cell type specific transcriptional changes, which could be potentially associated with disease pathogenesis and/or pathology. To investigate how disease state affected expression in a cell type specific manner, 94 samples from Blood CD4$^+$ T_{EM}, Blood CD19$^+$ B cells and LPL CD4$^+$ T_{EM} and CD19$^+$ B cells were sequenced on the Illumina Hiseq400 and Illumina HiSeq2500 platforms. Sequenced data were processed in collaboration with Medimmune Bioinformatics facility.

A total of DEG 688 genes were identified (Table 5.1), from which 5 included lncRNA. DEG with highest counts (and confidence) are summarized in Table 5.2.

Blood CD4$^+$ T_{EM} population displayed higher differences in expression profile between UC_n than UC_i when compared to controls. During the library generation only 3 Blood CD4$^+$ T_{EM} samples from UC_n passed quality checks, resulting in the smallest sample sub-population in our RNA study. To investigate this further a post-differential expression QC assessment were carried out. We noticed that an overwhelming majority of DEG were represented by very low gene counts. Unsurprisingly, we observed that large proportion of DEG that were associated with low counts also suffered from very high variance (For full details please see Appendix 3).

Table 5.1 NUMBER OF DEG IDENTIFIED IN SELECTED IMMUNE CELL POPULATIONS FROM HEALTHY SUBJECTS AND UC PATIENTS WITH DIFFERENT EXTENT OF DISEASE (*for n $_{donor}$ for each subset please see Table 3.2*). *Column headings show the cohorts that were compared, whereas population column shows the cell type compared. Direction and DEG columns describe: total number of DEG, DEG number upregulated or downregulated in UC $_{(i\ or\ n)}$ samples. LPL - Lamina propria; T_{EM} – T effector memory; UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon; UC(N) - Ulcerative colitis patient with non-inflamed Sigmoid colon; C - Control; ALL - total number of DEG; UP – number of upregulated DEG; DOWN - number of downregulated DEG.*

Control *vs* UC$_i$				**Control *vs* UC$_n$**		
Population	Direction	DEG		Population	Direction	DEG
LPL CD4$^+$ T_{EM} cells	ALL	291		LPL CD4$^+$ T_{EM} cells	ALL	63
	DOWN	121			DOWN	46
	UP	170			UP	17
LPL CD19$^+$ B cells	ALL	66		LPL CD19$^+$ B cells	ALL	17
	DOWN	37			DOWN	16
	UP	29			UP	1
Blood CD4$^+$ T_{EM} cells	ALL	22		Blood CD4$^+$ T_{EM} cells	ALL	157
	DOWN	2			DOWN	130
	UP	20			UP	27
Blood CD19$^+$ B cells	ALL	69		Blood CD19$^+$ B cells	ALL	3
	DOWN	22			DOWN	1
	UP	47			UP	2

Table 5.2 LIST OF DEG IDENTIFIED IN SELECTED IMMUNE CELL POPULATIONS FROM HEALTHY SUBJECTS AND UC PATIENTS WITH DIFFERENT EXTENT OF DISEASE. *Table shows all DEG that passed the filtering threshold of median count > 100. BaseMean represents mean normalized count over all samples in subgroup, p-adjusted shows p-value after corrected for multiple testing. LPL - Lamina propria; T_{EM} – T effector memory; UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon; UC(N) - Ulcerative colitis patient with non-inflamed Sigmoid colon; C – Control.*

Blood $CD4^+$ T_{EM} (Control *vs* UC_I)				
Symbols	**BaseMean**	**Log2Fold Change**	**p-adjusted**	**Gene Name**
ZNF585B	160.34	-1.85	1.16E-02	Zinc finger protein 585B
NBEAL2	847.27	0.87	5.70E-02	Neurobeachin like 2
CAD	238.89	1.26	9.73E-02	Carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase
FGL2	387.46	1.76	7.75E-02	Fibrinogen like 2
SIRPB1	111.57	2.53	7.36E-02	Signal regulatory protein beta 1
BCAS4	104.00	3.00	5.36E-02	Breast carcinoma amplified sequence 4
SRGAP1	105.23	3.10	3.94E-02	SLIT-ROBO rho gtpase activating protein 1

Blood $CD19^+$ B (Control *vs* UC_I)				
Symbols	**BaseMean**	**Log2Fold Change**	**p-adjusted**	**Gene Name**
RAPH1	109.75	-2.56	8.98E-02	Ras association (ralgds/AF-6) and pleckstrin homology domains 1
CEP44	106.69	-1.84	7.67E-02	Centrosomal protein 44
MIER3	338.51	-1.82	4.64E-02	MIER family member 3
ALMS1	544.48	-1.57	4.40E-02	ALMS1, centrosome and basal body associated protein
JUP	210.13	-1.54	9.67E-02	Junction plakoglobin
DYRK2	571.79	-1.13	5.81E-02	Dual specificity tyrosine phosphorylation regulated kinase 2
OPA1	570.69	-1.12	9.49E-03	OPA1, mitochondrial dynamin like gtpase
KLHL28	269.31	-1.12	5.81E-02	Kelch like family member 28
C10orf76	245.70	-1.06	9.82E-02	Chromosome 10 open reading frame 76
SLC39A10	507.57	-0.90	5.81E-02	Solute carrier family 39 member 10
STK17A	999.16	-0.84	9.49E-03	Serine/threonine kinase 17a
TMEM59	631.29	0.80	4.37E-02	Transmembrane protein 59
NCOA3	3384.15	0.81	1.96E-02	Nuclear receptor coactivator 3
CALR	5210.86	0.86	6.46E-02	Calreticulin
PDIA6	1505.30	0.91	8.81E-02	Protein disulfide isomerase family A member 6
HDLBP	1023.45	0.97	7.67E-02	High density lipoprotein binding protein
LDHA	921.57	0.98	8.25E-02	Lactate dehydrogenase A
RPN1	1850.38	1.01	5.81E-02	Ribophorin I
GAPDH	3454.42	1.02	9.49E-03	Glyceraldehyde-3-phosphate dehydrogenase
FAM46C	2837.80	1.07	6.38E-02	Family with sequence similarity 46 member C
CASP3	380.75	1.09	8.87E-02	Caspase 3
UBE2J1	2618.07	1.16	9.49E-03	Ubiquitin conjugating enzyme E2 J1
SEMA4A	184.67	1.18	5.43E-02	Semaphorin 4A
IDH2	409.54	1.18	2.70E-02	Isocitrate dehydrogenase (NADP(+)) 2, mitochondrial
NOMO1	521.75	1.19	6.38E-02	NODAL modulator 1
PPIB	2925.32	1.24	2.70E-02	Peptidylprolyl isomerase B
SRM	270.67	1.25	8.25E-02	Spermidine synthase
DENND1B	338.24	1.27	5.74E-02	DENN domain containing 1B
HSPA5	2117.76	1.27	2.70E-02	Heat shock protein family A (hsp70) member 5
ITM2C	1300.23	1.28	1.35E-02	Integral membrane protein 2C
PDIA4	2216.55	1.29	3.17E-02	Protein disulfide isomerase family A member 4
MAN1A1	1046.38	1.34	2.70E-02	Mannosidase alpha class 1A member 1
SAR1B	329.51	1.40	2.70E-02	Secretion associated ras related gtpase 1B
ARL14EP	314.74	1.44	1.96E-02	ADP ribosylation factor like gtpase 14 effector protein
ELL2	645.33	1.44	9.49E-03	Elongation factor for RNA polymerase II 2
HSP90B1	8423.62	1.46	7.30E-03	Heat shock protein 90 beta family member 1
HIST1H2AB	238.30	1.57	4.64E-02	Histone cluster 1 H2A family member b
BMP8B	179.69	1.59	4.40E-02	Bone morphogenetic protein 8b
FNDC3B	462.27	1.61	5.81E-02	Fibronectin type III domain containing 3B
TXNDC5	7027.65	1.66	3.23E-02	Thioredoxin domain containing 5
CD38	505.85	1.68	3.76E-02	CD38 molecule
MZB1	652.66	1.74	9.49E-03	Marginal zone B and B1 cell specific protein
GLDC	184.63	1.83	5.43E-02	Glycine decarboxylase
XBP1	2221.04	2.06	9.49E-03	X-box binding protein 1

LPL $CD4^+$ T_{EM} (Control *vs* UC_i)				
Symbols	**BaseMean**	**Log2Fold Change**	**p-adjusted**	**Gene Name**
ASAH2	114.11	-2.74	3.01E-02	N-acylsphingosine amidohydrolase 2
TTC7B	447.91	-2.44	5.57E-04	Tetratricopeptide repeat domain 7B
ZDHHC11B	144.69	-2.35	1.09E-02	Zinc finger dhhc-type containing 11B
OTUD3	236.88	-2.22	2.64E-02	OTU deubiquitinase 3
PBX1	129.71	-2.20	2.98E-02	PBX homeobox 1
SEMA4A	458.11	-2.16	6.82E-03	Semaphorin 4A
BAIAP3	102.71	-2.13	3.07E-02	BAI1 associated protein 3
EPB41L5	107.58	-2.01	9.62E-03	Erythrocyte membrane protein band 4.1 like 5
LATS2	118.53	-2.00	2.64E-02	Large tumor suppressor kinase 2
MATK	167.36	-1.99	3.17E-04	Megakaryocyte-associated tyrosine kinase
LMF1	121.00	-1.88	4.76E-03	Lipase maturation factor 1
UST	113.99	-1.82	7.61E-02	Uronyl 2-sulfotransferase
FOSB	14247.57	-1.78	2.23E-08	Fosb proto-oncogene, AP-1 transcription factor subunit
TEX101	421.29	-1.75	5.80E-02	Testis expressed 101
ARHGEF40	213.19	-1.73	3.96E-02	Rho guanine nucleotide exchange factor 40
CCDC88A	133.48	-1.68	4.40E-02	Coiled-coil domain containing 88A
LPAR6	168.75	-1.66	9.46E-02	Lysophosphatidic acid receptor 6
FOS	5124.11	-1.57	1.33E-07	Fos proto-oncogene, AP-1 transcription factor subunit
TPPP	221.97	-1.56	9.58E-03	Tubulin polymerization promoting protein
TANC2	317.66	-1.55	2.66E-02	Tetratricopeptide repeat, ankyrin repeat and coiled-coil containing 2
NAV1	224.26	-1.51	3.18E-02	Neuron navigator 1
SEMA4C	138.73	-1.50	6.74E-02	Semaphorin 4C
KIAA1683	947.73	-1.49	4.35E-03	Kiaa1683
DHRS3	209.33	-1.48	3.96E-03	Dehydrogenase/reductase 3
SORBS3	139.78	-1.45	3.77E-02	Sorbin and SH3 domain containing 3
LMNA	2404.14	-1.43	7.86E-04	Lamin A/C
C17orf107	252.25	-1.43	9.20E-03	Chromosome 17 open reading frame 107
AARS2	112.77	-1.40	6.35E-02	Alanyl-trna synthetase 2, mitochondrial
MYBBP1A	347.50	-1.38	7.12E-02	MYB binding protein 1a
NOL6	184.43	-1.35	5.42E-02	Nucleolar protein 6
SEC14L2	211.42	-1.34	9.23E-02	SEC14 like lipid binding 2
ADCY9	207.00	-1.34	6.87E-02	Adenylate cyclase 9
TYW3	117.15	-1.30	4.69E-02	Trna-yw synthesizing protein 3 homolog
ZNF417/ZNF587	198.65	-1.26	2.03E-02	Zinc finger protein 417
SGSM2	681.89	-1.25	2.04E-02	Small G protein signaling modulator 2
MYO9A	524.26	-1.25	7.55E-02	Myosin IXA
MAP1A	152.38	-1.25	9.33E-02	Microtubule associated protein 1A
GPR15	668.14	-1.25	4.56E-02	G protein-coupled receptor 15
TRIM39	459.05	-1.24	1.09E-02	Tripartite motif containing 39
ANXA1	2224.18	-1.23	7.86E-04	Annexin A1
THBS1	743.81	-1.21	6.87E-02	Thrombospondin 1
IGF1R	234.88	-1.20	3.80E-02	Insulin like growth factor 1 receptor
TTC9	175.85	-1.15	7.44E-02	Tetratricopeptide repeat domain 9
CCL5	1205.94	-1.14	8.44E-02	C-C motif chemokine ligand 5
NOL4L	562.43	-1.13	8.77E-02	Nucleolar protein 4 like
ZNF571	161.09	-1.12	9.23E-02	Zinc finger protein 571
CCDC125	323.50	-1.10	6.87E-02	Coiled-coil domain containing 125
TEDC1	372.20	-1.08	8.77E-02	Tubulin epsilon and delta complex 1
ZNF471	309.58	-1.07	5.94E-02	Zinc finger protein 471
LOC102724428/ SIK1	2166.74	-1.07	3.96E-03	Salt inducible kinase 1
IREB2	310.83	-1.06	5.97E-02	Iron responsive element binding protein 2
HIC1	293.37	-1.06	8.53E-02	HIC ZBTB transcriptional repressor 1
MED26	199.37	-1.02	5.20E-02	Mediator complex subunit 26
TXNIP	2621.52	-1.01	1.99E-02	Thioredoxin interacting protein
TIPARP	2282.72	-0.95	3.07E-02	TCDD inducible poly(adp-ribose) polymerase
WDR33	679.06	-0.94	9.23E-02	WD repeat domain 33
PRKCA	447.25	-0.94	2.04E-02	Protein kinase C alpha
MAB21L3	204.27	-0.92	8.75E-02	Mab-21 like 3
LRIG1	553.28	-0.88	9.29E-03	Leucine rich repeats and immunoglobulin like domains 1

NXF1	1044.99	-0.88	2.66E-02	Nuclear rna export factor 1
VIM	3368.08	-0.87	3.73E-02	Vimentin
FKBP4	435.21	-0.85	7.55E-02	Fk506 binding protein 4
DPYD	531.71	-0.84	9.28E-02	Dihydropyrimidine dehydrogenase
CASP8AP2	320.72	-0.84	9.49E-02	Caspase 8 associated protein 2
SORL1	2627.28	-0.84	1.34E-02	Sortilin related receptor 1
LARP4B	563.29	-0.74	9.49E-02	La ribonucleoprotein domain family member 4b
JUN	5955.00	-0.72	5.62E-02	Jun proto-oncogene, ap-1 transcription factor subunit
RNF216	571.49	-0.72	7.01E-02	Ring finger protein 216
PPP1R15A	6247.75	-0.69	3.42E-02	Protein phosphatase 1 regulatory subunit 15a
DUSP1	4010.09	-0.68	4.45E-02	Dual specificity phosphatase 1
B2M	14518.52	0.56	3.37E-02	Beta-2-microglobulin
CFLAR	2640.95	0.60	5.62E-02	Casp8 and fadd like apoptosis regulator
BIRC3	2270.93	0.67	9.23E-02	Baculoviral iap repeat containing 3
GAK	790.50	0.69	8.96E-02	Cyclin g associated kinase
CD2	2213.91	0.70	4.57E-02	Cd2 molecule
CALM3	1321.49	0.71	3.25E-02	Calmodulin 3
SURF4	791.48	0.73	9.07E-02	Surfeit 4
RANBP9	507.09	0.74	7.85E-02	Ran binding protein 9
ACTR3	1520.85	0.76	9.33E-02	Arp3 actin related protein 3 homolog
TPM4	1934.75	0.76	2.71E-02	Tropomyosin 4
RAP1B	992.04	0.76	3.07E-02	Rap1b, member of ras oncogene family
CBX3	484.36	0.77	3.07E-02	Chromobox 3
EFHD2	930.50	0.78	5.84E-02	Ef-hand domain family member d2
ZCCHC17	271.55	0.79	8.53E-02	Zinc finger cchc-type containing 17
CD247	1336.24	0.79	2.98E-02	Cd247 molecule
MAF	1786.21	0.80	7.21E-02	Maf bzip transcription factor
TAP1	886.67	0.82	2.23E-02	Transporter 1, atp binding cassette subfamily b member
SPOCK2	5596.82	0.83	3.21E-03	Sparc/osteonectin, cwcv and kazal like domains proteoglycan 2
SOD1	795.70	0.84	8.11E-02	Superoxide dismutase 1
KIF5B	1712.32	0.85	5.10E-02	Kinesin family member 5b
G3BP2	1779.13	0.86	3.37E-02	G3BP stress granule assembly factor 2
PKM	2068.32	0.89	3.03E-02	Pyruvate kinase M1/2
HMGB2	756.53	0.89	7.01E-02	High mobility group box 2
STARD7	609.00	0.90	8.32E-03	Star related lipid transfer domain containing 7
SLA	1095.40	0.91	9.91E-02	Src like adaptor
NAB1	798.96	0.92	1.77E-02	NGFI-A binding protein 1
GLCCI1	724.83	0.92	8.43E-02	Glucocorticoid induced 1
LDHA	1091.99	0.92	2.53E-02	Lactate dehydrogenase A
PIM3	400.63	0.93	3.03E-02	Pim-3 proto-oncogene, serine/threonine kinase
ADAM19	1514.53	0.95	4.58E-02	ADAM metallopeptidase domain 19
RNF19A	1641.17	0.95	2.05E-02	Ring finger protein 19A, RBR E3 ubiquitin protein ligase
PELI1	768.10	0.98	1.89E-02	Pellino E3 ubiquitin protein ligase 1
CD3D	1094.39	0.98	5.97E-02	Cd3d molecule
ICOS	1690.73	0.99	5.97E-02	Inducible t-cell costimulator
POLR3GL	398.42	0.99	4.21E-02	RNA polymerase III subunit G like
SREBF2	536.44	1.03	6.36E-02	Sterol regulatory element binding transcription factor 2
KAT2B	591.91	1.03	7.55E-02	Lysine acetyltransferase 2B
SRGN	1860.35	1.03	1.95E-02	Serglycin
BHLHE40	1560.97	1.07	7.19E-03	Basic helix-loop-helix family member e40
ST6GALNAC6	361.00	1.08	6.82E-02	ST6 n-acetylgalactosaminide alpha-2,6-sialyltransferase 6
DCP2	491.88	1.09	6.87E-02	Decapping mrna 2
TMEM173	280.01	1.10	2.44E-02	Transmembrane protein 173
FKBP5	1222.14	1.12	1.89E-02	FK506 binding protein 5
CD7	804.65	1.13	3.80E-02	CD7 molecule
BATF	309.07	1.13	3.13E-02	Basic leucine zipper atf-like transcription factor
L3MBTL3	428.12	1.13	2.66E-02	L3MBTL3, histone methyl-lysine binding protein
MICAL2	626.92	1.18	5.97E-02	Microtubule associated monooxygenase, calponin and LIM domain containing 2
PSMA4	343.03	1.18	3.36E-02	Proteasome subunit alpha 4
PSMB8	447.20	1.20	8.47E-04	Proteasome subunit beta 8
PTPRJ	1164.56	1.21	6.60E-04	Protein tyrosine phosphatase, receptor type J
IL12RB1	236.63	1.22	1.89E-02	Interleukin 12 receptor subunit beta 1

CTSC	665.95	1.22	3.13E-02	Cathepsin C
LYST	1234.06	1.23	3.75E-02	Lysosomal trafficking regulator
FAM53B	732.07	1.23	4.41E-04	Family with sequence similarity 53 member B
CLEC2B	327.04	1.23	2.66E-02	C-type lectin domain family 2 member B
GRAMD1B	398.16	1.26	1.96E-02	GRAM domain containing 1B
SEM1	302.26	1.26	1.09E-02	SEM1, 26S proteasome complex subunit
ZC3H7A	430.77	1.26	2.30E-03	Zinc finger ccch-type containing 7A
PRDM1	1326.67	1.28	1.09E-02	PR/SET domain 1
GBP5	1344.53	1.29	1.43E-02	Guanylate binding protein 5
TRAFD1	212.58	1.29	4.57E-02	Traf-type zinc finger domain containing 1
HIST1-H2AM	474.93	1.31	7.12E-02	Histone cluster 1 H2A family member m
DUSP16	1653.09	1.32	1.32E-03	Dual specificity phosphatase 16
ZNF282	200.75	1.35	9.51E-02	Zinc finger protein 282
IL21R	273.22	1.36	9.32E-02	Interleukin 21 receptor
FURIN	1610.52	1.44	2.21E-04	Furin, paired basic amino acid cleaving enzyme
IL1R1	310.46	1.44	7.19E-03	Interleukin 1 receptor type 1
GNA15	186.57	1.44	3.37E-02	G protein subunit alpha 15
CASS4	211.81	1.45	6.87E-02	Cas scaffolding protein family member 4
PTPN13	399.25	1.45	2.33E-02	Protein tyrosine phosphatase, non-receptor type 13
GBP2	1557.91	1.46	8.86E-05	Guanylate binding protein 2
HS3ST3B1	165.01	1.46	6.36E-02	Heparan sulfate-glucosamine 3-sulfotransferase 3B1
RORC	211.48	1.48	1.02E-02	RAR related orphan receptor C
TNFRSF18	212.55	1.52	1.32E-03	TNF receptor superfamily member 18
TNFRSF1B	1213.05	1.52	4.76E-03	TNF receptor superfamily member 1B
SMC2	168.97	1.54	4.25E-02	Structural maintenance of chromosomes 2
EZH2	223.58	1.55	7.84E-02	Enhancer of zeste 2 polycomb repressive complex 2 subunit
SYT11	223.49	1.56	4.86E-02	Synaptotagmin 11
TRIB2	687.58	1.57	3.99E-04	Tribbles pseudokinase 2
TRIM59	112.11	1.60	5.47E-02	Tripartite motif containing 59
CCNG2	192.44	1.61	1.33E-02	Cyclin G2
ZNRF1	172.09	1.63	8.82E-04	Zinc and ring finger 1
ICA1	120.77	1.67	9.91E-02	Islet cell autoantigen 1
FES	245.13	1.75	2.66E-02	FES proto-oncogene, tyrosine kinase
BCL2A1	120.21	1.76	2.66E-02	BCL2 related protein A1
MAST4	549.28	1.77	3.07E-03	Microtubule associated serine/threonine kinase family member 4
CTLA4	958.20	1.79	5.10E-03	Cytotoxic t-lymphocyte associated protein 4
NUSAP1	123.02	1.80	4.59E-02	Nucleolar and spindle associated protein 1
HIST2H3A	135.52	1.82	5.80E-02	Histone cluster 2 H3 family member a
TIFA	273.36	1.82	4.47E-03	TRAF interacting protein with forkhead associated domain
CCL20	157.27	1.86	3.25E-02	C-C motif chemokine ligand 20
HELLS	191.94	1.87	2.78E-02	Helicase, lymphoid specific
HIST1H2AI	284.62	1.88	4.43E-03	Histone cluster 1 H2A family member i
CXCR6	528.12	1.88	8.40E-06	C-X-C motif chemokine receptor 6
CSF1	359.01	1.89	3.23E-04	Colony stimulating factor 1
IL2RA	443.31	1.92	3.28E-06	Interleukin 2 receptor subunit alpha
NCAPG2	139.86	1.92	8.59E-02	Non-smc condensin II complex subunit G2
IKZF2	582.14	1.95	2.98E-03	IKAROS family zinc finger 2
HIST1H2AB	113.61	1.96	8.79E-02	Histone cluster 1 H2A family member b
DUSP4	1057.49	2.02	2.52E-08	Dual specificity phosphatase 4
SPATA17	100.02	2.02	9.07E-02	Spermatogenesis associated 17
LY75	273.04	2.04	1.91E-02	Lymphocyte antigen 75
ADTRP	203.54	2.10	9.65E-03	Androgen dependent TFPI regulating protein
HIST1H3C	201.77	2.17	8.98E-02	Histone cluster 1 H3 family member c
RCBTB1	154.01	2.18	3.04E-02	RCC1 and BTB domain containing protein 1
HIST1H3F	107.43	2.19	2.76E-02	Histone cluster 1 H3 family member f
HIST1H2AL	278.24	2.20	1.32E-03	Histone cluster 1 H2A family member l
ZEB2	700.46	2.22	3.23E-04	Zinc finger e-box binding homeobox 2
ENTPD1	1001.67	2.26	2.21E-04	Ectonucleoside triphosphate diphosphohydrolase 1
F5	212.75	2.27	5.82E-03	Coagulation factor V
HIST1H3B	299.23	2.43	9.49E-02	Histone cluster 1 H3 family member b
CD38	239.69	2.46	3.77E-05	CD38 molecule
HIST1H2AJ	632.01	2.53	3.42E-02	Histone cluster 1 H2A family member j
MKI67	299.76	2.56	6.96E-05	Marker of proliferation ki-67
MYO7A	156.07	2.60	9.62E-03	Myosin VIIA
PLEK	208.89	2.64	5.45E-03	Pleckstrin
MUC1	177.67	2.66	1.06E-02	Mucin 1, cell surface associated
TNIP3	301.27	2.82	1.24E-08	TNFAIP3 interacting protein 3
LAG3	578.86	2.83	8.22E-06	Lymphocyte activating 3
HAVCR2	113.43	3.00	1.24E-04	Hepatitis A virus cellular receptor 2
COL5A3	593.44	3.15	1.24E-08	Collagen type V alpha 3 chain
GNLY	248.98	3.63	9.49E-02	Granulysin
RRM2	144.97	3.68	3.55E-02	Ribonucleotide reductase regulatory subunit M2
DUOX2	919.09	5.99	4.66E-06	Dual oxidase 2
LCN2	285.64	6.38	4.35E-03	Lipocalin 2

LPL CD4$^+$ T_{EM} (Control *vs* UC_n)				
Symbols	**BaseMean**	**Log2Fold Change**	**p-adjusted**	**Gene Name**
KSR2	214.98	-2.93	2.69E-02	Kinase suppressor of ras 2
SPRED2	106.41	-2.90	4.47E-02	Sprouty related EVH1 domain containing 2
MTF1	272.34	-2.32	5.08E-02	Metal regulatory transcription factor 1
SNX27	276.01	-1.99	5.48E-03	Sorting nexin family member 27
GMCL1	138.65	-1.98	3.65E-02	Germ cell-less, spermatogenesis associated 1
ZNRF2	347.75	-1.79	5.03E-02	Zinc and ring finger 2
LMTK2	437.38	-1.73	8.05E-02	Lemur tyrosine kinase 2
ZNF587	198.65	-1.53	8.05E-02	Zinc finger protein 587
XPOT	380.03	-1.23	8.45E-02	Exportin for trna
YPEL3	276.85	-1.18	6.96E-02	Yippee like 3
PWP1	210.94	1.31	8.11E-02	PWP1 homolog, endonuclein
TYMP	293.36	1.52	5.46E-02	Thymidine phosphorylase
TMEM173	280.01	1.59	2.00E-02	Transmembrane protein 173
GPATCH11	204.20	1.70	8.45E-02	G-patch domain containing 11
C19orf24	102.52	2.00	4.89E-02	Chromosome 19 open reading frame 24
HELZ2	429.07	2.10	5.48E-03	Helicase with zinc finger 2
PMS1	121.91	2.18	4.46E-02	PMS1 homolog 1, mismatch repair system component
TNS3	171.98	2.26	6.38E-02	Tensin 3
COL6A1	153.21	2.70	8.11E-02	Collagen type VI alpha 1 chain

LPL CD19$^+$ B Cells (Control *vs* UC_i)				
Symbols	**BaseMean**	**Log2Fold Change**	**p-adjusted**	**Gene Name**
SMOC1	104.90	-2.63	4.61E-03	SPARC related modular calcium binding 1
CTSW	171.03	-2.50	7.98E-02	Cathepsin W
FRMD6	125.11	-2.26	8.38E-02	FERM domain containing 6
GPR107	338.55	-2.11	8.79E-02	G protein-coupled receptor 107
BMP6	371.02	-2.04	1.62E-03	Bone morphogenetic protein 6
CTSF	101.25	-1.89	8.24E-02	Cathepsin F
MIER3	275.20	-1.85	1.76E-02	MIER family member 3
WNT10A	166.12	-1.85	2.22E-02	Wnt family member 10A
GYG1	163.82	-1.76	4.73E-03	Glycogenin 1
TDRD3	130.00	-1.48	6.86E-03	Tudor domain containing 3
PAIP2B	281.79	-1.46	9.86E-03	Poly(a) binding protein interacting protein 2B
CRTAP	404.16	-1.43	3.99E-02	Cartilage associated protein
GPR15	1155.25	-1.40	3.07E-03	G protein-coupled receptor 15
C8orf33	164.15	-1.22	6.97E-02	Chromosome 8 open reading frame 33
PABPC4	2386.40	-1.12	4.36E-02	Poly(a) binding protein cytoplasmic 4
FOSB	11698.01	-0.93	8.79E-02	Fosb proto-oncogene, AP-1 transcription factor subunit
ST6GAL1	1732.53	0.77	6.22E-02	ST6 beta-galactoside alpha-2,6-sialyltransferase 1
MDM2	706.65	1.02	6.97E-02	MDM2 proto-oncogene
TAP1	452.56	1.15	9.86E-03	Transporter 1, ATP binding cassette subfamily B member
NOMO1	497.05	1.17	6.97E-02	NODAL modulator 1
PLD3	321.18	1.19	3.38E-02	Phospholipase D family member 3
PSME2	308.45	1.20	3.07E-03	Proteasome activator subunit 2
KLHDC4	137.33	1.24	7.00E-02	Kelch domain containing 4
FCGR2B	235.46	1.28	1.82E-02	Fc fragment of igg receptor iib
IFI30	405.04	1.48	1.82E-02	IFI30, lysosomal thiol reductase
CD38	491.78	1.50	1.82E-02	CD38 molecule
XPO4	232.11	1.74	6.84E-02	Exportin 4
ABCC10	145.27	1.76	7.00E-02	ATP binding cassette subfamily C member 10
DMBT1	102.58	5.33	6.84E-02	Deleted in malignant brain tumors 1
OLFM4	254.02	6.18	1.68E-02	Olfactomedin 4

LPL CD19$^+$ B Cells (Control *vs* UC_n)				
Symbols	**BaseMean**	**Log2Fold Change**	**p-adjustted**	**Gene Name**
ZYX	101.45	-2.05	3.85E-02	Zyxin
PLEKHA3	202.69	-2.85	4.98E-02	Pleckstrin homology domain containing A3
HIST2H3A	326.21	-3.55	4.98E-02	Histone cluster 2 H3 family member a
INTS1	332.12	-2.02	9.69E-02	Integrator complex subunit 1
MKI67	381.21	-3.14	6.04E-02	Marker of proliferation ki-67
HIST1H2AJ	1411.04	-3.65	9.03E-02	Histone cluster 1 H2A family member j

5.4.2 In-Silico Validation Of Genes Identified As Differentially Expressed Between The UC Patients And Healthy Individuals

Since we were not able to acquire additional samples for laboratory-based validation experiments (due the shortage of time), we decided to use computational methods. We would like to acknowledge that methods chosen are based on pre-existing expectations and provided with quick and superficial assessment of general data quality.

First, we manually screened the LPL $CD4^+$ $T_{EM\ (C\ vs\ UCi)}$ DEG lists for well-known and expected genes. LPL $CD4^+$ T_{EM} was selected as it is known to be a key cell population in IBD and it returned the highest number of differentially expressed genes. Previous studies in LPL $CD4^+$ T cells have showed elevated levels of IL17, RORC and IL23R production from patients with UC and proposed strong Th17 involvement in disease pathology (Kobayashi *et al.*, 2008). In addition, increased numbers of T_{reg} is another characteristics of active UC (Holmén *et al.*, 2006; Yu *et al.*, 2007).

We saw that, as expected, our LPL $CD4^+$ $T_{EM\ (C\ vs\ UCi)}$ DEG lists was enrichment for genes associated with Th17 signalling (↑*RORC,* ↑*IL17A,* ↑*IL17F,* ↑*IL1R1,* ↑*IL21,* ↑*IL21R)* and T_{reg} /T cell activation associated genes (↑*CTLA4,* ↑ *IL2RA,* ↑ *TNFRSF18,* ↑*CCL20*).

Second *in silico*-based method we used for data validation was an overlap assessment between newly identified DEG and already published transcriptomics studies. We selected one RNA Seq and one microarray-based study from PubMed literature search partially based on their expression data availability. *Van der Goten et al* used Affymetrix Human Gene 1.0ST array to detect the difference in transcriptional activity between control and UC_i, whereas *Taman et al* employed RNA Seq to look at expression differences between treatment-naïve UC patients and controls (Van der Goten *et al.*, 2014; Taman *et al.*, 2018).

Of the 65 DEG from LPL $CD19^+$ B cells $_{(C\ vs\ UCi)}$, 19% showed overlap with *Van der Goten et al* and 22% overlapped with *Taman et al.* Overlap for LPL $CD4^+$ T_{EM} was 20% and 26%, respectively. Finally, genes identified by *Van der Goten et al* and *Taman et al* were tested against each other and only 50% agreement was reached.

Next, we screened LPL $CD4^+$ $T_{EM\ (C\ vs\ UCi)}$ for genes which were identified as significantly different in all 3 studies. Genes that were shared and showed the same direction (↑UC) were associated with cell growth and proliferation (*RRM2, S100A9, MYBL2, IL21R, BHLHE40*) and T_{reg} signaling/T cell activation (*IL2RA, CTLA4, CD80, CCL20*). Interestingly, *DUOX2* showed strongest upregulation in UC_i (*Fold* $Change_{Taman}$ *= 53.08, Fold* $Change_{Goten}$ *= 32.87, Fold* $Change_{LPL\ CD4}$ *= 63.62)* than any other shared genes. Moreover, in our study *DUOX2* was associated with relatively high counts (median normalized counts = 919.09).

5.4.3 Determining Biological Meaning Behind The Disease Specific Change In Expression Profiles

To dissect biological processes and molecular pathways, enrichment analysis was performed using the annotations from GO and IPA. For these analyses, the higher the number of DEG called, the more informative the analysis, so enrichment was determined only for populations that contained more than 50 DEG. In addition, we used all genes, including ones with very low counts, reasoning that identification of molecular pathways might help to separate real signal from noise.

IPA analysis revealed that LPL $CD4^+$ T_{EM} cells $_{(C\ vs\ UCi)}$ were significantly enriched for T helper cell differentiation and T helper 2 pathways (Table 5.3). Interestingly Blood $CD19^+$ B cells $_{(C\ vs\ UCi)}$ where significantly enriched for endoplasmic reticulum stress pathway and unfolded protein response. Both observations were backed up by Gene Ontology, where LPL $CD4^+$ T_{EM} cells $_{(C\ vs\ UCi)}$ where enriched for immune and inflammatory response and Blood $CD19^+$ B cells $_{(C\ vs\ UCi)}$ for protein folding and ER-nucleus signalling pathways (Table 5.3).

As expected, DEG from Blood $CD4^+$ T_{EM} $_{(C\ vs\ UCn)}$ showed no-enrichment for any pathways. The initial differential expression calculation returned 150 genes, yet, no pathway-based clustering (± significant) was observed, strongly stating that despite large number of significant genes most of them are consequence of high noise.

Table 5.3 IPA PATHWAY ENRICHMENT ANALYSIS OF GENES DIFFERENTIALLY EXPRESSED IN BLOOD CD19$^+$ B CELL (CONTROL VS UCI), LPL CD19$^+$ B CELL (CONTROL VS UCN) AND LPL CD4$^+$ T_{EM} (CONTROL VS UCI) COHORTS. *Enrichment was calculated by Fishers-Exact test and adjusted for multiple testing. Pathway column reveals the pathway enriched for, -log(p-value) column shows negative log of p-value after adjustment for multiple testing, Ratio represents the proportion of all genes in pathway covered by DEG. Finally, DEG in Pathway displays the DEG which were enriched in pathway. LPL - Lamina propria; T_{EM} – T effector memory; UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon; UC(N) - Ulcerative colitis patient with non-inflamed Sigmoid colon; C – Control.*

Blood CD19$^+$ B cells (Control *vs* UC_I)			
Pathway	**-log(p-value)**	**Ratio**	**DEG in Pathway**
Endoplasmic Reticulum Stress Pathway	5.28	0.263	CALR,HSP90B1,CASP3,XBP1,HSPA5
Unfolded protein response	3.38	0.1	CALR,HSP90B1,PDIA6,XBP1,HSPA5

LPL CD19$^+$ B cells (Control *vs* UC_n)			
Pathway	**-log(p-value)**	**Ratio**	**DEG in Pathway**
Mitotic Roles of Polo-Like Kinase	1.55	0.0357	CCNB2,PLK1
Cell Cycle: G2/M DNA Damage Checkpoint Regulation	1.55	0.0426	CCNB2,PLK1

LPL CD4$^+$ T_{EM} (Control *vs* UC_i)			
Pathway	**-log(p-value)**	**Ratio**	**DEG in Pathway**
Differential Regulation of Cytokine Production in Intestinal Epithelial Cells by IL-17A and IL-17F	6.17	0.538	LCN2,CXCL1,CCL5,CSF2,CCL3,IL17F,IL17A
T Helper Cell Differentiation	5.84	0.204	IL21,HLA_DOA,CD80,IL12RB1,IL21R,ICOS,IL2RA, TNFRSF1B,RORC,IL17F,IL17A
Differential Regulation of Cytokine Production in Macrophages and T Helper Cells by IL-17A and IL-17F	5.51	0.545	CXCL1,CCL5,CCL3,CSF2,IL17F,IL17A
IL-17A Signalling in Gastric Cells	3.25	0.24	FOS,JUN,CCL20,CXCL1,CCL5,IL17A
Th2 Pathway	3.25	0.0968	CCR1,CD247,HLADOA,JUN,CD80,HAVCR1,IL12RB1,ICOS, MAF,CXCR6,IL2RA,CD3D
Th1 and Th2 Activation Pathway	3.2	0.0867	CD247,CCR1,HLA_DOA,JUN,CD80,HAVCR1,IL12RB1, HAVCR2,ICOS,MAF,CXCR6,IL2RA,CD3D
Role of Hypercytokinemia/hyperchemokinemia in the Pathogenesis of Influenza	3.1	0.294	CCR1,IL1RN,CCL5,CCL3,IL17A
Altered T Cell and B Cell Signalling in Rheumatoid Arthritis	3.01	0.136	IL21,HLA-DOA,CXCL13,CD80,CSF1,IL1RN,CSF2,IL17A
Role of IL-17A in Psoriasis	2.93	0.4	S100A9,CCL20,CXCL1,IL17A
Role of Cytokines in Mediating Communication between Immune Cells	2.76	0.238	IL21,IL1RN,CSF2,IL17F,IL17A
CCR5 Signalling in Macrophages	2.4	0.107	CD247,FOS,CALM1 (includes others),JUN,CCL5,CCL3,CD3D,PRKCA
Type I Diabetes Mellitus Signalling	2.4	0.0938	CD247,HLA_DOA,CD80,ICA1,GZMB,GAD1,IL1R1, TNFRSF1B,CD3D
Hematopoiesis from Pluripotent Stem Cells	2.19	0.25	CD247,CSF1,CSF2,CD3D
Glucocorticoid Receptor Signalling	2.11	0.0568	CD247,SELE,PBX1,CCL5,CCL3,CD3D,FOS,JUN,KAT2B, IL1RN,DUSP1,ANXA1,FKBP4,FKBP5,CSF2
HMGB1 Signalling	2.02	0.0796	FOS,SELE,JUN,KAT2B,IL1R1,CSF2,TNFRSF1B,IL17F, IL17A
Pathogenesis of Multiple Sclerosis	2.02	0.375	CCR1,CCL5,CCL3
Granulocyte Adhesion and Diapedesis	2.02	0.0796	SELE,CXCL13,IL1RN,CCL20,CXCL1,CCL5,IL1R1,CCL3, TNFRSF1B
Hepatic Cholestasis	1.99	0.0783	ADCY9,JUN,IL1RN,IL1R1,CSF2,TNFRSF1B,IL17F,PRKCA, IL17A
CD28 Signalling in T Helper Cells	1.99	0.0776	CD247,FOS,CALM1 (includes others), HLA-DOA,JUN,ACTR3,CD80,CD3D,CTLA4
Communication between Innate and Adaptive Immune Cells	1.99	0.115	B2M,CD80,IL1RN,CCL5,CCL3,CSF2
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	1.58	0.154	CD247,B2M,GZMB,CD3D
Cdc42 Signalling	1.58	0.0734	B2M,CD247,FOS,HLA-DOA,JUN,ACTR3,CD3D,IQGAP3
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	1.57	0.0512	SELE,CCL5,IL1R1,PDGFC,IL17A,CALM1 (includes others),FOS,JUN,IL1RN,CSF1,CSF2,TNFRSF1B,PRKCA
Nur77 Signalling in T Lymphocytes	1.57	0.111	CD247,CALM1 (includes others),HLA-DOA,CD80,CD3D
TNFR2 Signalling	1.57	0.148	FOS,JUN,TNFRSF1B,BIRC3
OX40 Signalling Pathway	1.57	0.111	B2M,CD247,HLA-DOA,JUN,CD3D
Agranulocyte Adhesion and Diapedesis	1.49	0.0684	SELE,CXCL13,IL1RN,CCL20,CXCL1,IL1R1,CCL5,CCL3
Graft-versus-Host Disease Signalling	1.44	0.133	HLA-DOA,CD80,IL1RN,GZMB
Calcium-induced T Lymphocyte Apoptosis	1.44	0.1	CD247,CALM1 (includes others),HLA-DOA,CD3D,PRKCA
Allograft Rejection Signalling	1.44	0.129	B2M,HLA-DOA,CD80,GZMB
IL-17A Signalling in Fibroblasts	1.44	0.129	FOS,JUN,LCN2,IL17A
Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes	1.42	0.0822	CD247,FOS,CALM1 (includes others),JUN,CD80,CD3D
Antigen Presentation Pathway	1.36	0.121	B2M,HLA-DOA,PSMB8,TAP1
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	1.34	0.0538	FOS,CALM1 (includes others),JUN,IL1RN,CSF1, IL1R1,CSF2,TNFRSF1B,BIRC3,IL17A

To further explore the functional role of genes identified as significantly different, manual PubMed search for each DEG (with median normalized count >100) was carried out. Table 5.4 shows the number of genes that had a PubMed hit when searched for IBD and/or UC.

From all differentially expressed genes in LPL $CD4^+$ $T_{EM\ (C\ vs\ UCi)}$ - *MKI67, CTLA4, CCL20, CD2, RORC, FOS, LCN2, MUC1, CCL5, CD38* were top 10 genes in terms of publication count when searched criteria was UC. *MKI67* is good representation of downfalls of term-based literature search. *MKI67* (also known as *Antigen KI-67*) is well established marker for cell proliferation and despite large number of publications returned, in most of cases, has no direct association with study itself. Other genes with high number of publications returned were:

- *CASP3* and *XBP1* for Blood $CD19^+$ B cells $_{(C\ vs\ UCi)}$;
- *MKI67* from LPL $CD19^+$ B cells $_{(C\ vs\ UCn)}$;
- *MDM2*, *CD38* and *DMBT1* for LPL $CD19^+$ B cells $_{(C\ vs\ UCi)}$;

We further looked for genes with less defined connection to UC (based on publication record). *TXNIP* (Thioredoxin Interacting Protein) encodes a protein of major importance in redox balance (Spindel, World and Berk, 2012) and was detected as significantly decreased in LPL $CD4^+$ T_{EM} cell in UC patients with active inflammation. *Takahashi et al 2007* showed that colonic tissue from UC patients had reduced *TXNIP* expression in comparison to healthy controls. However, they mainly attributed the decrease in *TXNIP* levels due loss of topical epithelial cell layer in inflamed tissue (Takahashi *et al.*, 2007).

SEMA4A (Semaphorin 4A) returned only two publications in PubMed search for its role in IBD. SEMA4A encodes a protein with broad functionality, including T-cell mediated immune response (Ito and Kumanogoh, 2016). In contradiction to *Vadasz et al 2015, who reported* increase in *SEMA4A* levels in patients with both - UC and CD (Vadasz *et al.*, 2015), we observed significant decrease in *SEMA4A* transcript levels in LPL $CD4^+$ T_{EM} from UC_i.

Another, interesting gene was *KSR2* (Kinase Suppressor of Ras 2) - here identified as significantly downregulated in LPL CD4$^+$ T_{EM} from UC_n. Even thought, KSR2 itself has no associations with either UC or IBD, *Xue et al 2013* showed that micro RNA miR-31 (which is significantly dysregulated in UC (Gwiggner *et al.*, 2018), downregulates KSR2 and facilitates IL-2 secretion and T cell activation (Xue *et al.*, 2013).

Finally, in addition to genes with already established functions and/or associations, substantial part of DEG had no known functional role at all. Such genes included histone proteins (*HIST1H2AB, HIST1H2AI, HIST1H2AJ, HIST1H2AL, HIST1H2AM, HIST1H3B, HIST1H3C, HIST1H3F, HIST2H3A)* and zinc finger proteins (*ZNF282, ZNF417/ZNF587, ZNF471, ZNF571, ZNRF1*) which were identified as differentially expressed between UC_i and control in LPL CD4$^+$ T_{EM} cells.

Table 5.4 PUBMED LITERATURE SEARCH FOR DEG IDENTIFIED IN STUDY. *Table shows the number of DEG for each cell population which were used for search and number of hits. Hit represents a gene which returned at least one publication associated with either UC or IBD. Hit table shows the symbol for all genes that had any previous record. LPL - Lamina propria; T_{EM} – T effector memory; UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon; UC(N) - Ulcerative colitis patient with non-inflamed Sigmoid colon; C – Control.*

Population	Number DEG Tested	Number of Hits	Hits
LPL CD4$^+$ T_{EM} cells (C *vs* UC_i)	**195**	**80**	ASAH2; OTUD3; SEMA4A; FOSB; FOS; LMNA; ADCY9; GPR15; TRIM39; ANXA1; THBS1; IGF1R; IGF1R; CCL5; SIK1; HIC1; TXNIP; PRKCA; NXF1; VIM; FKBP4; JUN; RNF216; PPP1R15A; DUSP1; B2M; CFLAR; BIRC3; GAK; CD2; CALM3; TPM4; CBX3; MAF; TAP1; SOD1; HMGB2; SLA; GLCCI1; ADAM19; CD3D; ICOS; SREBF2; KAT2B; TMEM173; FKBP5; CD7; BATF; PSMB8; PTPRJ; IL12RB1; PRDM1; IL21R; IL1R1; RORC; TNFRSF18; TNFRSF1B; EZH2; TRIB2; ICA1; FES; CTLA4; CCL20; CXCR6; CSF1; IL2RA; DUSP4; LY75; ZEB2; ENTPD1; F5; CD38; MKI67; PLEK; MUC1; LAG3; HAVCR2; GNLY; DUOX2; LCN2;
LPL CD4$^+$ T_{EM} cells (C *vs* UC_n)	**19**	**5**	SPRED2; TYMP; TMEM173; PMS1; COL6A1;
Blood CD4$^+$ T_{EM} cells (C *vs* UC_i)	**7**	**2**	CAD; FGL2;
LPL CD19$^+$ B cells (C *vs* UC_i)	**30**	**14**	CTSW; BMP6; WNT10A; GPR15; FOSB; ST6GAL1; MDM2; TAP1; PSME2; FCGR2B; IFI30; CD38; DMBT1; OLFM4;
LPL CD19$^+$ B cells (C *vs* UC_n)	**6**	**1**	MKI67;
Blood CD19$^+$ B cells (C *vs* UC_i)	**45**	**15**	RAPH1; JUP; TMEM59; PDIA6; GAPDH; CASP3; SEMA4A; IDH2; SRM; DENND1B; HSPA5; PDIA4; HSP90B1; CD38; XBP1;

5.4.4 Determining Anatomical Location Specific Change In Expression Profiles In Purified CD19$^+$ B Cells And CD4$^+$ T_{EM} Immune Cell Populations

We next attempted to establish how anatomical location might affect the transcriptional landscape for cells of the same sub-type. The cell type was defined by the same surface marker expression. Transcription profiles of blood and SC LPL resident CD4$^+$ T_{EM} and CD19$^+$ B cells from healthy individuals were compared. The same pipeline as for disease specific differential expression profiling was used.

A total of 520 and 839 genes were identified as significantly different between blood and SC LPL residing CD4$^+$ T_{EM} and CD19$^+$ B cells. A substantial proportion of DEG from CD19$^+$ B Cells $_{(Blood\ vs\ LPL)}$ were associated with low counts and, non-surprisingly, high log2fold change. DEG between Blood and LPL CD4$^+$ T_{EM} cells had higher median normalized counts than CD19$^+$ B cells. The top 10 genes from both comparisons are summarized in Table 5.5.

Biological pathway identification by IPA showed that differentially expressed genes between Blood and LPL in CD19$^+$ B cells were enriched for unfolded protein response (*p = 3.84e-05* ; *ATF4; CALR; CANX; DDIT3; DNAJB9; DNAJC3; HSP90B1; HSPA5; HSPA1A/HSPA1B*; OS9; P4HB; PDIA6; PPARG; PPP1R15A; SEL1L; XBP1*), PI3K signaling in B lymphocytes (*p = 7.57e-05* ; *AKT3; ATF3; ATF4; ATF5; BLK; CARD10; CD180; CD79B; FCGR2B; FOS; FOXO3; IL4R; ITPR1; ITPR3; JUN; LYN; NFKBIA; PDIA3; PIK3AP1; PLCB4; PLCL2; PLEKHA2; PRKCB; PTPRC*) and antigen presentation pathway (*p =7.57e-05* ; *CALR; CANX; CD74; CIITA; HLA-DMB; HLA-DOB; HLA-DPA1; HLA-DBP1; HLA-DQB1; HLA-DRA; HLA-DRB1; PDIA3*) (Table 5.6).

Significantly different genes from CD4$^+$ T_{EM} $_{(Blood\ vs\ LPL)}$ was enriched for protein kinase A signalling (*p = 3.62e-05* ; *ADCY6; AKAP9; ATF4; CALM1; CDC25B; CREM; DUSP1; DUSP2; DUSP4; DUSP8; HIST1H1C; HIST1H1D; HIST1H1E; ITPR3; LEF1; MAP3K1; MYL12A; NFKB2; NFKBIA; PDE4A; PDE4B; PDED; PLCB3; PLCD3; PPP1R10*) and integrin linked kinase pathway (*p = 2.83e-03* ; *ATF4; CDH1; DOCK1; DSP; FOS; IRS2; ITGB1; ITGB4; JUN; KRT18; LEF1; MYC; NACA; NFKB2; PDGFC; RND3; TMSB10/TMSB4X; VEGFA*) (Table 5.6).

Table 5.5 LIST OF TOP 10 DEG (BY LOG2FOLDCHANGE) IDENTIFIED IN SELECTED IMMUNE CELL POPULATIONS FROM PERIPHERAL BLOOD AND SIGMOID COLON LAMINA PROPRIA

Table shows all DEG that passed the filtering threshold of median count > 100. BaseMean represents mean normalized count over all samples in subgroup, p-adjusted shows p-value after corrected for multiple testing. LPL - Lamina propria; T_{EM} – T effector memory.

CD19$^+$ B Cells (Blood *vs* LPL)				
Symbols	**BaseMean**	**Log2Fold Change**	**p-adjusted**	**Gene Name**
ZNF860	121.76	-5.58	3.07E-04	Zinc finger protein 860
FCN1	285.16	-5	3.48E-02	Ficolin 1
C1orf162	103.19	-4.47	8.24E-03	Chromosome 1 open reading frame 162
PCDH9	362.46	-4.38	7.01E-05	Protocadherin 9
SELL	1595.69	-4.12	3.53E-06	Selectin L
LGALS4	349.12	10.45	2.21E-09	Galectin 4
KIT	160.94	10.62	9.83E-05	KIT proto-oncogene receptor tyrosine kinase
CLCA1	287.43	11.32	1.66E-08	Chloride channel accessory 1
FABP1	439.03	11.53	6.99E-09	Fatty acid binding protein 1
KRT20	409.32	12.18	2.32E-09	Keratin 20

CD4$^+$ T_{EM} (Blood *vs* LPL)				
Symbols	**BaseMean**	**Log2Fold Change**	**p-adjusted**	**Gene Name**
HBB	102.77	-8.4	6.72E-02	Hemoglobin subunit beta
S1PR1	1066.69	-4.36	1.77E-03	Sphingosine-1-phosphate receptor 1
PLEK	390.04	-3.95	7.84E-03	Pleckstrin
EOMES	102.06	-3.39	5.80E-02	Eomesodermin
CCDC65	118.16	-2.95	2.03E-02	Coiled-coil domain containing 65
PIGR	3869.55	10.69	1.19E-09	Polymeric immunoglobulin receptor
CEACAM7	202.06	11.25	2.41E-03	Carcinoembryonic antigen related cell adhesion molecule 7
KRT20	310.64	11.84	2.39E-04	Keratin 20
CLCA1	366.93	12.03	1.49E-03	Chloride channel accessory 1
FABP1	344.45	12.04	6.36E-06	Fatty acid binding protein 1

Table 5.6 IPA PATHWAY ENRICHMENT ANALYSIS OF GENES DIFFERENTIALLY EXPRESSED IN CD19 B CELL (BLOOD *vs* LPL) AND $CD4^+$ T_{EM} (BLOOD *vs* LPL) COHORTS. *Enrichment was calculated by Fishers-Exact test and adjusted for multiple testing. Pathway column reveals the pathway enriched for, -log(p-value) column shows negative log of p-value after adjustment for multiple testing, Ratio represents the proportion of all genes in pathway covered by DEG. Finally, DEG in Pathway displays the DEG which were enriched in pathway. LPL - Lamina propria; T_{EM} – T effector memory.*

$CD4^+$ T_{EM} (Blood *vs* LPL)			
Pathway	**-log(p-value)**	**Ratio**	**DEG in Pathway**
EIF2 Signalling	52.4	0.387	RPL11,RPL22,RPS27,RPS23,RPL35A,HSPA5,RPS11,MYC, RPS7,VEGFA,RPL13,RPS20,RPS13,RPL23A,RPL19,RPL21 ,ATF4,IRS2,RPL36,RPL32,ATF3,RPL36AL,RPL3,RPS8, RPL29,RPL12,EIF3E,RPL37A,RPL10A,RPL15,RPS6,RPL8, HNRNPA1,PPP1R15A,RPS25,RPS15A,RPL6,RPL41, RPL13A,RPSA,RPL24,RPS3A,RPS18,RPL26,RPL7A,RPL7, RPL27A,RPL14,RPL35,RPL18A,UBA52,RPS9,RPS5,RPS12 ,RPS3,RPS17,RPL31,RPL4,EIF3H,RPL34,RPL17,RPL30, RPS10,RPL23,RPS21,RPS29,FAU,RPL27,RPS15,RPS16, RPS27A,RPL5,RPL37,RPL38,RPS14
mTOR Signalling	11.4	0.189	RPS3A,RPS27,RPS18,RPS23,PDGFC,RPS11,VEGFA, RPS7,RPS20,RPS13,RPS9,IRS2,RPS17,RPS3,RPS12, RPS5,EIF3H,DDIT4,RPS8,RPS10,RPS21,EIF3E,RPS29 ,FAU,RPS6,RPS15,RPS16,RND3,RPS27A,RPS15A,RPS25, RPS14,RPSA
Regulation of eIF4 and p70S6K Signalling	11.4	0.208	RPS3A,RPS27,RPS18,RPS23,RPS11,RPS7,RPS20,RPS13, RPS9,IRS2,RPS5,RPS17,RPS12,RPS3,ITGB1,EIF3H,RPS8, RPS10,RPS21,EIF3E,RPS29,FAU,RPS6,RPS15,RPS16, RPS27A,RPS25,RPS15A,RPSA,RPS14
Protein Kinase A Signaling	4.44	0.104	HIST1H1C,DUSP8,AKAP9,PDE4A,DUSP2,PTPRF,CDC25B, PLCD3,NFKBIA,PPP1R10,ATF4,PDE4D,MYL12A,PTPRD, PPP1R1B,MAP3K1,ADCY6,NFKB2,PDE4B,TTN,CALM1 (includes others),TULP2,PTPRH,HIST1H1E,DUSP1,CREM, ITPR3,PLCB3,HIST1H1D,LEF1,DUSP4,SFN
ILK Signalling	2.55	0.115	ITGB1,NFKB2,PDGFC,MYC,VEGFA,FOS,DOCK1,CDH1, JUN,RND3,ATF4,IRS2,LEF1,KRT18,ITGB4, TMSB10/TMSB4X,DSP,NACA
TNFR2 Signaling	2.45	0.259	FOS,JUN,NFKBIA,MAP3K1,TNFAIP3,NFKB2,TRAF1
Toll-like Receptor Signaling	1.77	0.158	FOS,JUN,NFKBIA,UBA52,MAP3K1,TNFAIP3,RPS27A, NFKB2,TRAF1
cAMP-mediated signalling	1.66	0.104	ADRA2B,RGS2,AKAP9,PDE4A,ADCY6,PDE4B,CALM1 (includes others),TULP2,DUSP1,CREM,S1PR1,ATF4, PDE4D,DUSP4,PTGER4

CD19^{+} B cells (Blood *vs* LPL)			
Pathway	**-log(p-value)**	**Ratio**	**DEG in Pathway**
Unfolded protein response	4.42	0.314	PPARG,CALR,P4HB,DDIT3,HSPA1A/HSPA1B,XBP1,CANX, DNAJC3,OS9,DNAJB9,HSPA5,SEL1L,HSP90B1,PDIA6, PPP1R15A,ATF4
Hepatic Fibrosis / Hepatic Stellate Cell Activation	4.12	0.209	CXCL8,COL19A1,IL4R,COL5A2,ICAM1,COL4A1,COL6A2, KLF6,FGFR2,COL4A2,IFNAR2,VEGFA,COL1A2,COL5A1, TGFBR2,COL16A1,COL1A1,COL6A1,IGF1,COL13A1,A2M, EGFR,COL3A1
Protein Kinase A Signalling	4.12	0.144	ENPP6,PDIA3,PDE3A,PTPRF,DUSP2,TGFBR2,PTPRC, DUSP5,NFKBIA,PPP1R10,DUSP26,PTPRO,DUSP10,ATF4, SMPDL3B,GNG12,MYL12A,KDELR1,PTPN6,PTPRD, YWHAE,PDE9A,YWHAB,PPP1R1B,MAP3K1,ADCY6,PYGB, ITPR1,PLCL2,PTPRH,PLCB4,HIST1H1E,H3F3A/H3F3B, ADD3,DUSP1,ITPR3,PTPRS,HIST1H1D,IHH,DUSP4,H1F0, PTPN21,KDELR2,PRKCB
PI3K Signaling in B Lymphocytes	4.12	0.207	IL4R,ATF3,CD79B,ATF5,PDIA3,PLCL2,ITPR1,FCGR2B, PTPRC,BLK,FOS,PLCB4,JUN,NFKBIA,CD180,CARD10, ITPR3,FOXO3,LYN,AKT3,ATF4,PIK3AP1,PLEKHA2, PRKCB
Antigen Presentation Pathway	4.12	0.353	CALR,HLA-DRB1,PDIA3,HLA-DMB,HLA-DRA,CIITA,CANX,HLA-DOB,CD74,HLA-DQB1,HLA-DPB1,HLA-DPA1
GP6 Signalling Pathway	3.75	0.206	COL19A1,COL5A2,COL4A1,COL6A2,FGFR2,ITPR1, COL4A2,COL16A1,COL5A1,COL1A2,FGFR3,COL1A1, COL6A1,COL13A1,LAMB3,LAMA3,LYN,AKT3,COL3A1, RASGRP2,PRKCB
Th2 Pathway	3.21	0.189	SOCS3,IL4R,ICAM1,TNFRSF4,IKZF1,FGFR2, HLA-DQB1,IL24,TGFBR2,FGFR3,ITGB2,JUN, HLA-DRB1,HLA-DRA,GFI1,HLA-DMB,S1PR1 ,HLA-DOB,HLA-DPB1,ACVR1C,HLA-DPA1
Aldosterone Signalling in Epithelial Cells	2.29	0.158	PDIA3,SGK1,HSPA1A/HSPA1B,DNAJC3,FGFR2,DNAJC1, DNAJC25,PLCL2,ITPR1,DNAJB9,HSPA5,FGFR3,SCNN1A, HSP90B1,PLCB4,DUSP1,DNAJB11,ITPR3,HSPA13, HSP90AA1,DNAJB1,PRKCB
Th1 and Th2 Activation Pathway	2.29	0.159	SOCS3,IL4R,ICAM1,TNFRSF4,IKZF1,FGFR2, HLA-DQB1,IL24,TGFBR2,FGFR3,ITGB2,JUN,HLA-DRB1,HLA-DRA,GFI1,HLA-DMB,S1PR1,HLA-DOB,HLA-DPB1,DLL4,ACVR1C,HLA-DPA1
Endoplasmic Reticulum Stress Pathway	2.25	0.368	CALR,HSP90B1,DDIT3,XBP1,ATF4,DNAJC3,HSPA5
B Cell Development	2.05	0.333	PTPRC,HLA-DRB1,CD79B,HLA-DMB,HLA-DRA,HLA-DOB,HLA-DQB1
Role of Tissue Factor in Cancer	2.05	0.172	CXCL8,P4HB,EGR1,ITGA6,PLAUR,FGFR2,F3,BLK,VEGFA, FGFR3,YES1,PAK1,F2RL1,PDIA6,LYN,AKT3,EGFR
Osteoarthritis Pathway	2.05	0.152	PPARG,CXCL8,EPAS1,RARRES2,DDIT4,BMP2,ITLN1, SMAD6,WNT16,ACVRL1,SLC39A8,FGFR3,VEGFA,TGFBR2,ELF3,DDR2,RUNX2,FOXO3,IHH,ATF4,SOX9
Dendritic Cell Maturation	2.04	0.151	ICAM1,PDIA3,LTB,FGFR2,CD83,PLCL2,HLA-DQB1,FCGR2B,COL1A2,FGFR3,COL1A1,PLCB4,NFKBIA, HLA-DRB1,DDR2,HLA-DRA,HLA-DMB,AKT3,HLA-DOB,ATF4,COL3A1
CD28 Signalling in T Helper Cells	2.03	0.162	PTPN6,MAP3K1,FGFR2,ITPR1,HLA-DQB1,PTPRC,FGFR3,FOS,PAK1,HLA-DRB1,NFKBIA,JUN,HLA-DRA,ITPR3,HLA-DMB,AKT3,ARPC3,HLA-DOB
OX40 Signalling Pathway	1.82	0.222	JUN,HLA-DRB1,TNFRSF4,NFKBIA,HLA-DMB,HLA-DRA,HLA-DOB,HLA-DQB1,HLA-DPB1,HLA-DPA1
Granulocyte Adhesion and Diapedesis	1.55	0.173	ITGB2,HRH1,CXCL8,SELL,CLDN23,SDC1,ICAM1,PF4, MMP15,ITGA6,CLDN7,CXCL2,CLDN3
Calcium-induced T Lymphocyte Apoptosis	1.5	0.2	HLA-DRB1,ITPR3,HLA-DMB,HLA-DRA,NR4A1,HLA-DOB,HLA-DQB1,ITPR1,ATP2A2,PRKCB
Allograft Rejection Signaling	1.39	0.25	HLA-DRB1,HLA-DMB,HLA-DRA,HLA-DOB,HLA-DQB1,HLA-DPB1,HLA-DPA1
CXCR4 Signaling	1.3	0.137	EGR1,ADCY6,FGFR2,ITPR1,BCAR1,FGFR3,FOS,DOCK1, PLCB4,PAK1,JUN,RND3,ITPR3,LYN,AKT3,GNG12,MYL12A ,PRKCB

5.5 Discussion

Here, we successfully purified and sequenced $CD4^+$ T_{EM} cells and $CD19^+$ B cells from peripheral blood and LP of SC colon. This allowed us, for first time, to dissect the transcriptional differences between healthy and UC at cell type specific level. In total we identified 688 and 1359 genes significantly affected by disease state and origin, respectively.

Apart from Blood $CD4^+$ T_{EM} populations, patients with active inflammation had greater differences in their expression profiles than UC patients with no inflammation, when compared to control. Though it is expected for patients with active disease to bear larger differences, the interpretation is challenged by the fact that the recruitment rate for UC_n were markedly reduced when compared to UC_i. Thus, lower rates of observed DEG may also reflect differences in experimental power, as well as genuine biology.

Rigorous pre- and post- differential expression QC allowed us to identify that a large proportion of DEG were called based upon very low on gene counts. Moreover, when combined with low sample numbers, as in case for Blood $CD4^+$ T_{EM} UC_n, this proved insufficient for reliable assessment of underlying biological difference in expression profiles. Power calculations showed that our current data set can be used to identify prognostic genes that are associated with high average counts (hereby lower dispersion) and fold change. However, further validation with qPCR or other expression-based method, such as, RNA scope, is crucial but was not possible at current time frame.

Instead to build a confidence in our results, while keeping in mind the time limitations, we screened our newly discovered data sets for already known and published markers. We confirmed that our LPL $CD4^+$ T_{EM} $_{(C\ vs\ UCi)}$ DEG lists had increased expression for genes associated with Th17 signalling. Yet, only around 1/5 of genes identified between LPL $CD4^+$ T_{EM} $_{(C\ vs\ UCi)}$ or LPL $CD19^+$ B cells $_{(C\ vs\ UCi)}$ showed an overlap to DEG from other already published expression studies. We speculated that the fact that both selected studies used whole tissue instead of purified lymphocytes might be one of the

main factors behind low overlap observed. Other discrepancies between or study and published data were heterogeneity in biopsy collection site (Van der Goten *et al.*, 2014) and disease activity (Taman *et al.*, 2018).

Finally, we proceeded to investigate the biological meaning behind the disease specific change in expression profiles. We showed that DEG between LPL CD4$^+$ T_{EM} cells $_{(C\ vs\ UCi)}$ were significantly enriched for T_H2 pathway ($p = 1x10^{-3.25}$). For a long time, UC has been strongly associated with T_H2 response. Early *in vitro* stimulation of UC derived LPL T cells showed increased expression of IL-5 (Fuss *et al.*, 1996) and, later, IL-13 (Fuss *et al.*, 2004), both known to be markers for T_H2 cells. Since then, a substantial number of papers has been published and even drug trials, targeting T_H2 associated cytokine – IL-13, carried out. The results of these studies has been rather contradicting, with one camp showing reduced secretion of IL-13 in UC (Vainer *et al.*, 2000; Kadivar *et al.*, 2004; Biancheri *et al.*, 2014) and questioning the importance of T_H2 pathway in UC. The additional fuel to T_H2 importance in UC came from observation that there was no significantly beneficial effect by IL-13 blockade (Danese *et al.*, 2015; Reinisch *et al.*, 2015). Expression of Interleukin-4 (IL-4) - another signature cytokine produced during the T_H2 response – is low and does not change in response to disease (NIESSNER and VOLK, 2008). Meanwhile, other labs showed increased *IL-13, IL-5* expression in active UC when compared to UC_n (Reinisch *et al.*, 2015), inactive UC (Inoue *et al.*, 1999) and control (Nemeth *et al.*, 2017).

Even though we showed significant enrichment for T_H2 pathway *($p = 1x10^{-3.25}$)*, neither of the abovementioned T_H2 signature cytokines nor two main TF - *GATA3* and *STAT6* - showed differential expression in LPL CD4$^+$ T_{EM} (C *vs* UC_i). Instead the enrichment was based on chemokine receptor (↑*CXCR6,* ↑*CCR1*), IL receptor (↑*IL12RB1,* ↑*IL2RA*) and TCR subunits (↑*CD3D,* ↑*CD247*), T_H2 associated TF (↓*JUN,* ↑*MAF*), co-stimulator (↑*ICOS*) and even APC associated proteins (↓*HLA-DOA,* ↑*CD80*) and hepatitis virus cellular receptor (↑*HAVCR1*). Moreover, five of these DEG enriched for T_H2 pathway were also associated with T_H differentiation *($p = 1x10^{-5.84}$)*. Hereby, we feel that we lack the evidence to reliably infer evidence of increased T_H2 activity status in our UC_i samples.

Both HLA-DOA and CD80 are expressed by APCs and should not be present in our T_{EM} data set. However, for pathway analysis we used all DEG, including DEG associated with very low counts. Indeed, after closer look we saw that both HLA-DOA and CD80 and HAVCR1 were far below our threshold for reliable expression and most likely represent the sequencing noise. However, we would like to acknowledge that when looking at the population purity based on negative sort marker expression, possible contamination was noticed.

Th17 activity was supported by enrichment for both - T_H Differentiation and Differential Regulation of Cytokine production in Macrophages and T Helper Cells by IL-17A and IL-17F pathways. Even though, there was no difference in ether *STAT3* nor *IL23R* expression, Th17 associated TF (↑*RORC*), secreted IL (↑*IL-17A,* ↑*IL-17F* and ↑*IL-21*) and IL17A target molecules (↑*CCL3,* ↑*CXCL1,* and ↑*CSF2*) where enriched amongst the DEG. With exception to RORC, all other genes associated with Th17 function were below the counts threshold. However, due the larger number of DEG present, we believe that increase in Th17 activity in UC_i is real. Our observation is in agreement with other published studies showing increased IL-17 expression by intestinal T cells isolated from UC patients (Fujino *et al.*, 2003; Kobayashi *et al.*, 2008).

In conclusion, our study identifies disease state specific changes in transcriptional landscape in purified $CD19^+$ B cells and $CD4^+$ T_{EM} immune cell populations from peripheral blood and sigmoid colon. Due to low gene counts (high variance) and small number of samples results should be interpreted with caution. In addition, the initial QC metrics looking at the expression of negative sort markers showed that there might be some cross-contamination. Therefore, to increase confidence in current findings, further validation is essential.

All wet laboratory-based validation experiment still faces problem of limited material availability, possibly in suboptimal quality. Thus, to validate all DEG lists single cell sequencing has the highest probability to yield good results. However, for only a few interesting hits, verification based upon qPCR might be method of choice. For genes with large expression differences, RNA scope might be more beneficial as it does not

require any nucleic acid extraction or manipulation; instead whole tissue can be stained for as many as 12 RNA species allowing easy visual comparison.

Expression comparison of other published studies would be an alternative dry-lab based method for data validation. With support from large expression studies, such as Human Cell Atlas, single cell data from healthy and UC patients from blood and colon have become available (unfortunately only after our experimental work was finished) and would be the best data set to compare our findings.

6. Comparison Of Chromatin Accessibility Between The Healthy Volunteers And UC Patients

6.1 Introduction

Measurement of chromatin organization permits the capture of the physical accessibility of regulatory elements at individual cell type resolution (Lee *et al.*, 2004; John *et al.*, 2011; Thurman *et al.*, 2012). *Degner et al 2012* showed that single genetic variants can alter both the chromatin conformation and magnitude of gene expression. In addition, they showed that these genetic variants are enriched in TFB sites (Degner *et al.*, 2012), and thus possibly reflects the allele-specific effects on TF binding.

We hypothesized that IBD risk associated variants could contribute to disease development by altering the function of regulatory elements, such as TFB sites, which in turn would be reflected in chromatin accessibility. Indeed, the observation that some of the SNPs fine mapped in IBD are enriched for transcriptional factor binding sites and tissue specific epigenetic marks further supports our hypothesis (Huang *et al.*, 2017).

In this chapter, we performed chromatin profiling of purified cell populations from peripheral blood and Lamina propria and Intraepithelial layers of the SC. We attempted to estimate the differential openness between similar cell types taken from healthy and diseased bowel.

6.2 Aim

- To decipher the chromatin landscape in purified cell populations from healthy volunteers and UC patients.
- To investigate differences in chromatin accessibility on a cell type specific and anatomical region-specific basis in states of disease and health.

6.3 Materials And Methods

6.3.1 ATAC Seq Library Sequencing

All ATAC Seq libraries were sequenced at Wellcome Trust Sanger Institute, Hinxton. Illumina HiSeq 2500 instruments with v4 chemistry was used for sequencing, with 10 samples per line. 75bp PE reads with 5% Phix spiking was selected.

6.3.2 ATAC Seq Data Analysis Pipeline

Initial analysis on raw ATAC Seq data was performed by Dr. J. Gutierrez-Achurry (post-doctoral research associate in Dr Carl Anderson's group, Wellcome Trust Sanger Institute), whereas further downstream analysis was carried out by the author.

6.3.2.1 Pre-Processing Of Raw Sequencing Data

The initial analysis in a stepwise manner:

- Raw ATAC Seq read QC by *FastQC* quality metrics tool;
- Adapter trimming and short read (<25nt) removal;
- Second QC by *FastQC* quality metrics tool;
- Trimmed read alignment to human reference genome (hg38) using *BWA* (Li and Durbin, 2009);
- Removal of mitochondrial and duplicate reads and ENCODE black listed regions by *Bedtools* (Quinlan and Hall, 2010) and *Picard;*
- Sequencing quality evaluation based on alignment using *Bedtools;*
- Peak calling with *MACS2* algorithm (*--nomodel --shift -25 - - extsize 50*) (Zhang *et al.*, 2008);

The final result of this initial pipeline was peak coordinates and refined read files accompanied by a set of QC reports from each stage of initial analysis.

6.3.2.2 Downstream Analysis

6.3.2.2.1 Peak Consolidation

The term “peak” is used for an identified region of open chromatin, identified by an excess of sequencing reads relative to a threshold determined by the background rate observed in any given region of chromatin.

Use of the MACS2 algorithm allowed *de-novo* identification of peaks in each individual sample, yet careful consideration is needed to separate the peaks representing underlying biology from the individual noise. It is very intricate process as in a given phenotype, peak might be present only in fraction of samples within a group. In addition, additional technical challenges, such as differences in sequencing depth, introduce inter-sample variation. The most straight forward way would be to select all peaks that are open in at least one sample per phenotype, yet it would result in excessive list of peaks which would lead to massive burden on multiple testing and introduction of false positives. In contrast, setting stringent filtering criteria will lead to loss of true positives. Moreover, MACS2 identified peaks have unique coordinates in each of the samples, making peak validation and uniform coordinate establishment more complex.

Unfortunately, at time there was no ATAC Seq specific analysis pipelines written that can tackle these challenges. Instead, in line with common practice, we adapted algorithms generated for other sequencing data analysis. To this end, we selected DiffBind (Stark and Brown, 2011) and DESeq2 (Love, Huber and Anders, 2014) R packages and modified their associated workflows to fit our dataset. DiffBind provides the user with a flexible peak filtering algorithm where the user can manually specify the percentage of samples in which a peak must be present to classify the chromatin as unfolded.

In our case, we considered a region truly open if a peak was present in at least 50% of samples for any given group, where the genomic coordinates for that peak was set to include the entire region covered by all peaks in individual samples that overlapped by at least 1 base pair (Figure 6.1)

We grouped samples based on the cell type and disease state they represented. For each of group we generated a consolidated peakset containing peaks with highest likelihood of representing underlying biology.

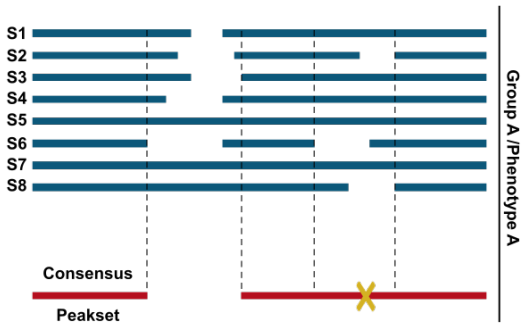


Figure 6.1. GRAPHICAL ILLUSTRATION OF STEPS INVOLVED IN PEAK FILTERING AND UNIFIED COORDINATE ESTABLISHMENT. *In this example a region is considered open if >50% from total of 8 samples have an open region that overlaps by at least 1 base pair. The consensus peakset is a name given to the object that contains genomic coordinates for all peaks which passed filtering criteria.*

6.3.2.2.2 Peak Annotation

ChipSeeker R package (Yu, Wang and He, 2015) was used to assign annotations to all peaks present.

Before annotation all peak-containing objects were filtered to exclude both sex chromosomes (e.g. chrX and chrY). Next, Transcription start site (TSS) region was set to ±1Kb from TSS and *TxDb.Hsapiens.UCSC.hg38.knownGene* transcript level annotations (Bioconductor Core Team, 2019) used to assign the genomic features and genes intersecting or closest to peaks. Following priority was adopted for genomic features:

1. Promoters
2. 5’ UTRs
3. 3’UTR,
4. Exon,
5. Intron,
6. Downstream
7. Distal Intergenic

In addition, we used *ChipSeeker* to obtain list of genes flanking (±5Kb) each peak.

6.3.2.2.3 Counts Matrix Generation Based On Peak Coordinates In Consolidated Peaksets

As next step in our analysis we generated counts matrices based on the regions classified as truly open. To define a region as differentially accessible, a comparison of normalized counts between populations of interest is carried out. Establishment of a unified coordinate system between all samples to be compared is crucial as read counts are dependent on peak width.

Consolidated peaksets contained peaks and their associated coordinates unique for disease state, anatomical location and cell type. We further combined these peaksets to create new objects that would contain all peak coordinates necessary for either exploratory analysis or differential accessibility calculation.

After construction of objects containing unified peak coordinates, the *dba.count* function (from DiffBind) was evoked. It counts how many reads overlap each interval for each unique sample. The result of counting was a counts matrix in DiffBind called *binding affinity matrix.* We named the count matrix which represented all peaks in study *Global binding affinity matrix.* We used the *Global binding affinity matrix* to assess the sample reproducibility.

The count matrix for the differential accessibility calculation were called individual *binding affinity matrix.* A total of 10 individual *binding affinity matrices* were constructed and included open regions from*:*

- SC Epithelium UC_i and UC_n and C
- SC IEL $CD4^+$ T_{EM} UC_i and UC_n and C
- SC IEL $CD8^+$ T_{EM} UC_i and UC_n and C
- SC LPL $CD4^+$ T_{EM} UC_i and UC_n and C
- SC LPL $CD8^+$ T_{EM} UC_i and UC_n and C
- SC LPL $CD19^+$ B cells UC_i and UC_n and C
- Blood $CD4^+$ T_{EM} UC_i and UC_n and C
- Blood $CD8^+$ T_{EM} UC_i and UC_n and C
- Blood $CD19^+$ B cells UC_i and UC_n and C
- Blood $CD14^+$ MF UC_i and UC_n and C

6.3.2.2.4 Sample Selection And Data Quality Control

The same as for RNA seq data, we performed an extensive ATAC seq data QC. For in depth details of QC and major challenges encountered during ATAC seq analysis please see Appendix 6 and Appendix 8.

6.3.2.2.5 Call For Differential Accessibility

For initial attempts to investigate differences in chromatin opening between control and UC patients, the DiffBind package was used. DiffBind allows a choice of method by which the differential accessibility analysis will be calculated, and in this case, we selected the same model as DESeq2 use. To assess if DESeq2 function inherit data normalization method would be suitable for ATAC seq data series of simulation experiments was performed (For more detail please see Appendix 7).

After selecting the median ratio method as the most suitable for our ATAC Seq normalization, all individual *binding affinity matrices,* to be used for differential accessibility calculation, were filtered to remove all peak regions with geometric mean of normalized read count below 30. Finally, significance was calculated by *DESeq2* algorithm.

After calling for differential accessibility in this manner, the call performance was evaluated using a range of quality control metrics (For more detail please see Appendix 6).

For second attempt to determine differences in chromatin conformation *DiffBind class objects* were transformed into DESeq2 specific objects. Each counts matrix was filtered so that peaks would be retained only if all samples had more than 10 counts each. *design* was set on condition (e.g. UC_i, UC_n and C) and analysis for differential accessibility repeated.

6.4 Results

6.4.1 Determining Chromatin Landscape In Purified Cell Populations From Peripheral Blood And Sigmoid Colon Lamina Propria And Intraepithelial Layers From Healthy Volunteers And UC Patients

Our first aim was to characterise chromatin landscape in terms of open and closed regions. We used ATAC seq technique to profile the chromatin structure in 10 different purified cell populations, including, $CD4^+$ or $CD8^+$ T_{EM} cells from blood, SC LP and SC IEL, $CD19^+$ B cells from blood and SC LP, $CD14^+$ MF from blood and $CD326^+$ epithelial cells from SC LP.

We identified a total of 393931 accessible regions present in at least 1 of 30 phenotypes, where each individual phenotype was unique for either disease state or anatomical location or cell type (Figure 6.2).

Next, we asked how many of open regions observed in a cell type were shared between different disease states. If two peaks overlapped by at least 1bp they were classified as shared. We observed that in average 90.77 $\pm$ 9.82%, 64.51 $\pm$ 11.74% and 55.76 $\pm$ 9.72% of peaks identified in any given Control, UC_i and UC_n conditions were shared. Cell types from healthy controls had less unique peaks than the other two categories. However, it is important to highlight that number of accessible regions were in moderate negative correlation with the number of samples for any given cell type (*r = -0.517, p = 0.0034*) as well as moderate positive correlation with sequencing depth (*r = 0.495, p = 0.0054*) (Figure 6.3).

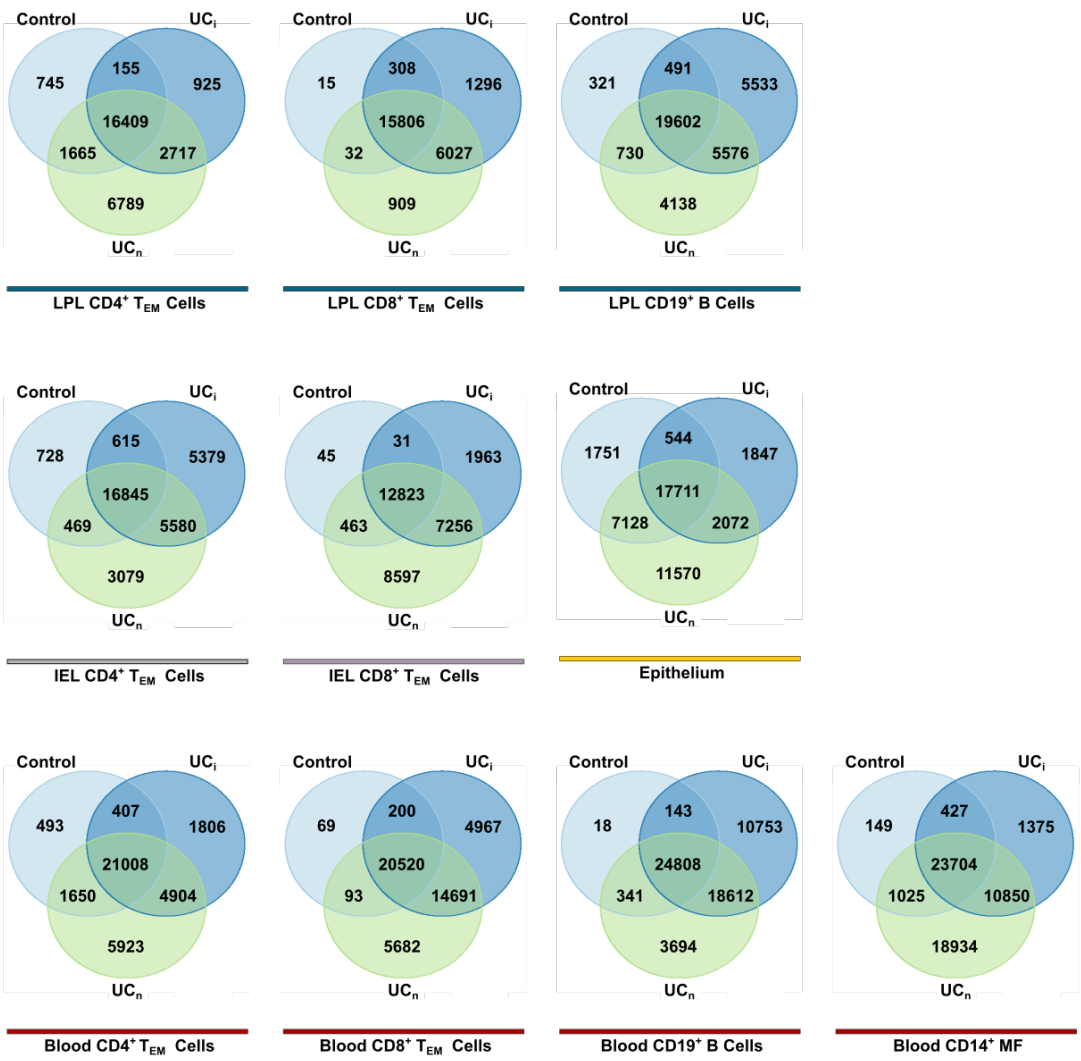


Figure 6.2 VENN DIAGRAMS SHOWING OPEN CHROMATIN REGIONS THAT WERE UNIQUE TO OR SHARED BETWEEN DIFFERENT DISEASE STATES IN A CELL TYPE AND ANATOMICAL LOCATION SPECIFIC MANNER (*n $_{phenotype}$ = 30; For n $_{donor}$ please see table 3.2*). *Light blue circles represent the Control samples, dark blue UC$_i$ and green shows the peak distribution from UC$_n$. LPL - Lamina propria; IEL – Intraepithelial lymphocytes; T$_{EM}$ – T effector memory; UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon; UC(N) - Ulcerative colitis patient with non-inflamed Sigmoid colon.*

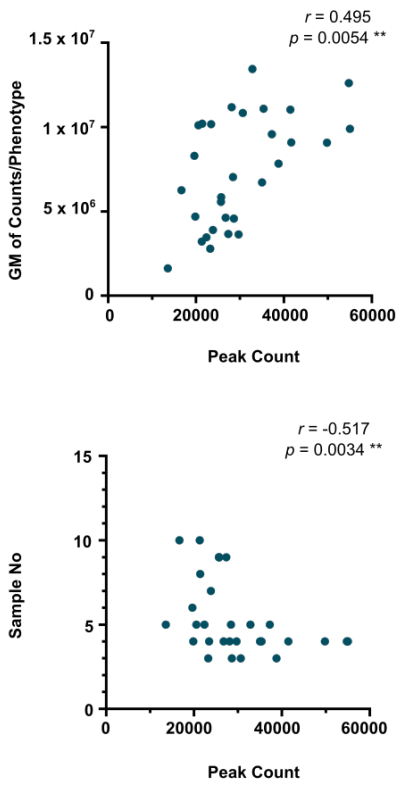


Figure 6.3 SCATTER PLOTS ILLUSTRATING THE RELATIONSHIP BETWEEN THE PEAK COUNT, SAMPLE NUMBER AND GEOMETRIC MEAN OF COUNTS ASSOCIATED WITH THE SAME PHENOTYPE ($n_{phenotype} = 30$)**.** *All relationship was quantified by Spearman's correlation. r - Spearman's rho; p - p-value; Sample No – Sample number; GM – Geometric mean.*

To gain better biological interpretation, all peaks were annotated to genomic features, such as, Promoter, 5UTR, 3UTR, Exon, Intron, Downstream, Intergenic, they showed overlap with. To accomplish this task ChipSeeker annotation package (Yu, Wang and He, 2015) was used.

The highest percentage of peaks mapped to the promoter regions (18% - 60%), followed by intronic (9.5% -36.5%) and distal regions (9% - 26%). Only small percentage of peaks mapped to the 3'UTRs, exonic and Downstream regions (Figure 6.4).

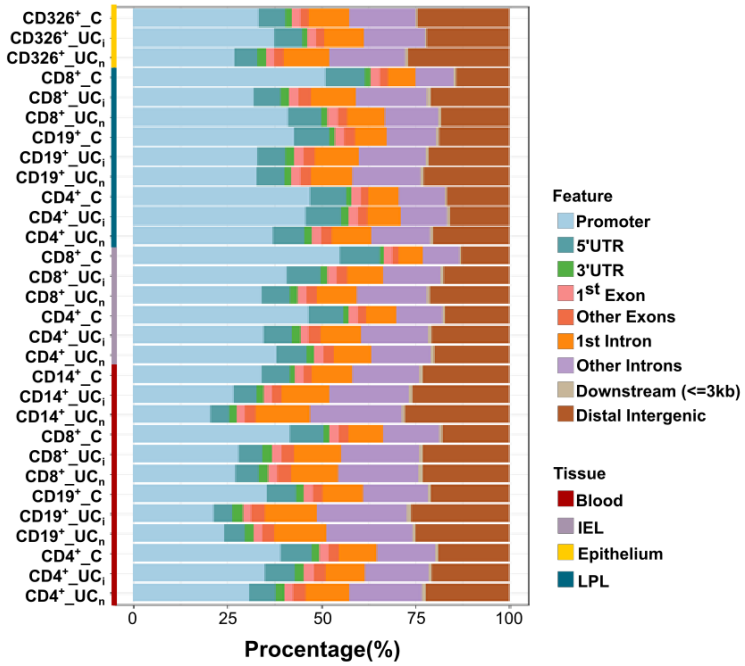


Figure 6.4 GENOMIC ALIGNMENT OF ACCESSIBLE REGIONS. *Legend shows the tissue types and genomic features, and their associated colour. LPL - Lamina propria; IEL – Intraepithelial lymphocytes; UC_i - Ulcerative colitis patient with inflamed Sigmoid colon; UC_n - Ulcerative colitis patient with non-inflamed Sigmoid colon; C – Control.*

6.4.2 Determining Differences In Chromatin Accessibility In Purified Cell Populations From Peripheral Blood And Sigmoid Colon Lamina Propria And Intraepithelial Layers From Healthy Volunteers And UC Patients

After successful identification of open chromatin regions in 30 different phenotypes of interest, we were intrigued to see if we could quantify how different disease states affects chromatin accessibility in a cell type specific manner.

We acknowledge that we have very small sample size per each phenotype and that any outcome of our analysis must be validated before any reliable conclusions can be made. We proceeded with analysis hoping to potentially find few interesting regions that we could follow up in less time consuming and expensive biological assays.

For first call of differential accessibility, we used DESeq2 model integrated into DiffBind analysis package. During QC we found that in majority of comparisons test statistics had failed. We performed series of troubleshooting experiments (summarized in Appendix 6) and repeated the differential accessibility analysis, this time using DESeq2 package. More comparisons now passed our QC pipeline (Table 6.1). However, the large number of DAR identified between healthy and UC_n in Blood $CD8^+$ T_{EM} and $CD19^+$ B cells seemed unusual, and, hereby, all DA calculations should be treated with caution.

Table 6.1. LIST OF DIFFERENTIALLY ACCESSIBLE REGIONS. *First column shows the cell type, second the conditions compare; third and fourth column displays the number DAR upregulated and downregulated. Finally, the last column summarizes the total DAR count identified in comparison. LPL - Lamina propria; IEL – Intraepithelial lymphocytes; T_{EM} – T effector memory; INF - Ulcerative colitis patient with inflamed Sigmoid colon; NON INF - Ulcerative colitis patient with non-inflamed Sigmoid colon.*

		UP	DOWN	TOTAL
Blood $CD19^+$ B cells	INF vs Control	2934	3087	6021
	NON. INF vs Control	2911	3190	5801
Blood $CD4^+$ T_{EM} cells	INF vs Control	148	212	360
	NON. INF vs Control	1	0	1
Blood $CD8^+$ T_{EM} cells	INF vs Control	1001	1129	2130
	NON. INF vs Control	118	959	1077
LPL $CD19^+$ B cells	INF vs Control	3084	3538	6622
	NON. INF vs Control	0	1	1
LPL $CD8^+$ T_{EM} cells	INF vs Control	9	62	71
	NON. INF vs Control	5	4	9
LPL $CD4^+$ T_{EM} cells	INF vs Control	443	758	1201
	NON. INF vs Control	9	26	35
IEL $CD8^+$ T_{EM} cells	INF vs Control	87	359	446
	NON. INF vs Control	12	30	42
Epithelium	INF vs Control	2	6	8
	NON. INF vs Control	0	2	2

6.5 Discussion

Growing evidence has shown that chromatin accessibility is itself under genetic control and can result in changes in gene expression. Taken together with observation that GWAS loci are enriched with chromatin quantitative treat loci, this makes chromatin architecture an attractive target for understanding the molecular mechanisms underlying the UC risk. Here, we used ATAC Seq to profile the chromatin landscape in primary immune cells and epithelium from peripheral blood and SC from healthy and diseased.

Together we identified thousands of accessible regions and mapped them to genomic features and genes they overlie or where closest to. We showed that chromatin profile at individual population level had the highest alignment score to promoter regions. Our finding goes against what other chromatin conformation studies have reported. First, a whole genome chromatin conformation study showed that only 16% of all DNase I hypersensitive sites were located on first exon or within 2kB upstream of promoter region with highest mapping rates on introns (39%) and intragenic regions (39%) (Boyle *et al.*, 2008). The low TSS region mapping was further supported by Thurman et al 2012, where only 3% (n = 75,575) of DHSs localize to TSS. However, *Boyle et al* after evaluating a variety of genome annotation resources concluded that at the time when these earlier experiments were carried out annotations were not complete and completely reliable yet. We used the latest human transcriptome annotation resource, which included TSS for protein coding and non-coding genes. We feel that with addition of protein non-coding genes the number of peaks mapping to TSS regions might have grown.

We speculated that the low percentage of TSS reads reported by *Thurman et al* was consequence of analysis method employed by the researchers. The genomic annotations were assigned to an object which contained all peaks present in 130 cell types profiled. At the same paper they showed that peaks on promoters are significantly more conserved between different cell types than any other genomic regions. Hereby,

we question in the small percentage of reads mapping to promoter regions are consequence of high conservation of this chromatin region.

Next, we attempted to quantify if and how much does disease state and inflammation change the chromatin profile. Unfortunately, we found that for majority of comparisons test statistics either failed or produced data that passed all QC but still were at suboptimal quality.

We could not identify the exact reason behind poor performance of test statistics. However, during corrections we had a meeting with Dr Blagoje Soskic (Postdoctoral Fellow at Wellcome Sanger Institute) who has considerable experience with working with ATAC seq data. During our discussion it was shared that their team had experienced the same failure in p-value distribution. They too used DESeq2 with small replicate number (3-4). They had not looked for exact reason behind it, but under-sequencing was mentioned as one of possible reason. Following the work performed in Dr Dan's Gaffney's lab, it seems that 300M reads per sample could be the appropriate sequencing depth to perform the analysis of interest.

Interestingly, very recent publication by *Gontarz et al., 2020* showed that in comparison to other methods used for differential accessibility estimation, such as limma and edgeR, DESeq2 performed very poorly when sample size dropped below 7 and data were associated with low average counts (*Gontarz et al., 2020*).

In summary, we identified open and close chromatin regions in 10 different cell populations from peripheral blood and sigmoid colon lamina propria and intraepithelial layers from healthy volunteers and UC patients. Our current sampling and sequencing strategy and possibly data analysis algorithm were too weak to allow us to determine if there is significant change in chromatin accessibility between different disease states with confidence. However, if we had more time, there are ways we could strengthen our current data sets, which could then allow us to perform differential accessibility analysis. One of the methods would be to go back and physically recruit more participants while doing much deeper sequencing. The second method, would be combining our data with already published ATAC Seq data. However, we currently

are not aware of any human data sets assessing the chromatin landscape in gut residing immune cells. Thought, for both - blood $CD4^+$ T cells and B cells, chromatin profiles have been generated and could be used to further assess and refine our data.

7. Combinatorial Analysis Of Functional Genomics And Genetics Data

7.1 Introduction

In chapters 5 and 6 we attempted to employ a "phenotype-first" approach and identify changes in the gene expression and chromatin conformation between healthy control and UC patients in a cell type specific manner. We assumed that results from these analysis pipelines are directly related to the disease, although this association might be either causal or simply reflect downstream changes of underlying inflammation. However, neither expression nor chromatin profiling alone provides with a comprehensive view of potential regulatory processes implicated in disease development.

As final part of my PhD we aimed to integrate GWAS results with functional genomics data generated during my PhD. We hypothesized that this approach would allow us to gain a comprehensive mechanistic interpretation of how known disease associated SNPs affect UC and potentially resolve causality by linking observed differences to underlying genetic elements (which by definition can be causal but never an effect of inflammation).

Unfortunately, during the data analysis we discovered that both ATAC seq and RNA seq data are of poor quality, mostly due to the very shallow sequencing depth. Moreover, when combined with low sample numbers as for ATAC seq experiments - gives an incomplete survey of chromatin accessibility and, hereby, cannot be used to determine important disease biology. However, we decided to proceed with the analysis as learning exercise to improve authors computational and data analyses skill set.

7.2 Aims

- Determine the extent by which difference in transcriptional profiles between the control and diseased tissue are reflected in differences in chromatin conformation and vice versa.
- Assess if differentially expressed genes and differentially accessible regions are enriched for genetic variants associated with autoimmune diseases or treats.
- Combine GWAS, ATAC Seq and RNA Seq data to propose the causal links by which variants contained in risk loci might regulates gene expression with respect to specific genes and specific cell types.

7.3 Materials And Methods

7.3.1 RNA Seq Data Correlation With ATAC Seq Data

Lists of all genes and peaks that were used for differential expression or accessibility calculations and passed DESeq2 filtering were collected. Transcriptomics data represent differences in expression of protein coding genes only. Low expressed genes were associated with high noise in our data set. Therefore, we used an R script to remove all genes with mean expression value below 100 counts.

Next, ATAC Seq peak lists were passed to *ChipSeeker* for annotation to overlapping genomic feature, such as, promoter, 5'UTR, 3'UTR, intron, exon, distal or downstream region, based on peak location and gene they overlapped or were closest to. It is known that open promoter regions in some degree reflect active gene expression (Boyle et al., 2008), yet there is less evidence about how other open genomic feature, such as peaks distal to a TSS, may impact on transcriptional activity for nearby genes. Therefore, we calculated correlation for each region separately.

Some of the genes showed overlap with multiple peaks. There is no established method to model the way in which multiple peaks might influence expression of a single gene and different previous studies have used variety of methods. Due to its simplicity, we decided to partially follow model published and used by *Scott-Browne et al 2016*. In short, when a gene had multiple peaks, each peak was assigned to the same gene with the same transcriptomic expression data, but each peak retained its own original estimate of relative openness based upon log2fold change ATAC Seq data.

The data matrix with gene-peak pairs and their associated relative expression and accessibility values in terms of Log2FoldChange was transferred to Prism Graph Pad software and correlation calculated by Spearman Rho. The relationship between the expression and accessibility was assessed, so that:

- DEG – DA (only significantly different genes and peaks located on genomic feature of interest, such as promoter, would be matched)

- DA - All genes (significantly different peaks with all genes expressed by the same populations)
- All peaks - DEG (all peaks located on genomic feature of interest with significantly different genes)
- All peaks - All genes (all peaks located on genomic feature of interest with all genes expressed by the same samples)

7.3.2 Calculation Of Genes And Chromatin Regions Enrichment Within GWAS Risk Loci

Enrichment of DEG and DA within the GWAS risk loci associated with immune mediated diseases and traits was calculated using the algorithm written by Dr Tim Raine (Raine *et al.*, 2015). The modified algorithm, encoded for *python* (v2 and v3), is summarized in Figure 7.1.

Lists of all genes and peaks used for differential expression and openness calculation with their associated relative expression values in terms of Log2Foldchange and p-value were exported as data matrices. Next, we used *biomaRT* R package to retrieve the genomic coordinates for each gene and peak mapping to Human reference genome v37 (Hg37). Gene and peak coordinates were backwards converted to a previous human genome build (GRCh37), as focal SNPs coordinates were mapped to this previous build. In instances where regions or genes could not be converted to this build, they were excluded from further analysis.

Lists of genetic risk loci associated with immune disease, such as, UC (de Lange *et al.*, 2017a), CD (de Lange *et al.*, 2017a), Type 1 Diabetes (T1D) (Barrett *et al.*, 2009), Rheumatoid Arthritis (RhArt) (Okada *et al.*, 2014), Coeliac (International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), 2015) and traits - Mean Corpus Volume (MCV) (Astle *et al.*, 2016), Body Mass Index (BMI) (Locke *et al.*, 2015), and Height (Wood *et al.*, 2014)- were identified from studies deposited in the National Human Genome Research Institute - European Bioinformatic Institute (NHGRI – EBI) published GWAS catalogue (NHGRI-EBI, 2017). Only variants with p – value $\leq 5x10^{-8}$ was exported.

The SNP list was further filtered to remove all variants falling on sex chromosomes and the region around the MHC.

The recombination map was downloaded from PLINK command-line program (Chang, 2019) accompanying web resources. The *python* script, written for version 2, was run from Unix command line.

All association lists were pre-filtered to remove duplicate loci keeping just focal SNPs with best association statistics within any given LD block. This approach does reduce the number of SNPs associated with any given locus (and might miss instances where multiple independent SNPs are associated with disease – e.g. as with NOD2 locus) but allows for pooling of multiple different GWAS without inappropriate duplication of replicated loci. In order to account for nonrandom patterns of chromatin conformation extra caution was made when peak background was modeled.

Next, an interval extending 0.1cM on either side of each focal SNP was defined. The frequency of DEG or DA falling within these set regions was determined and compared to a reference/null distribution defined by overlap testing of GWAS SNPs with repeated random samples of equivalent numbers of genes or open peaks drawn from the same cell population without restriction to those showing differential expression or differential openness. For expression and chromatin data, DEG and DA lists were further subdivided by genes/peaks up or downregulated in UC and tested for enrichment separately.

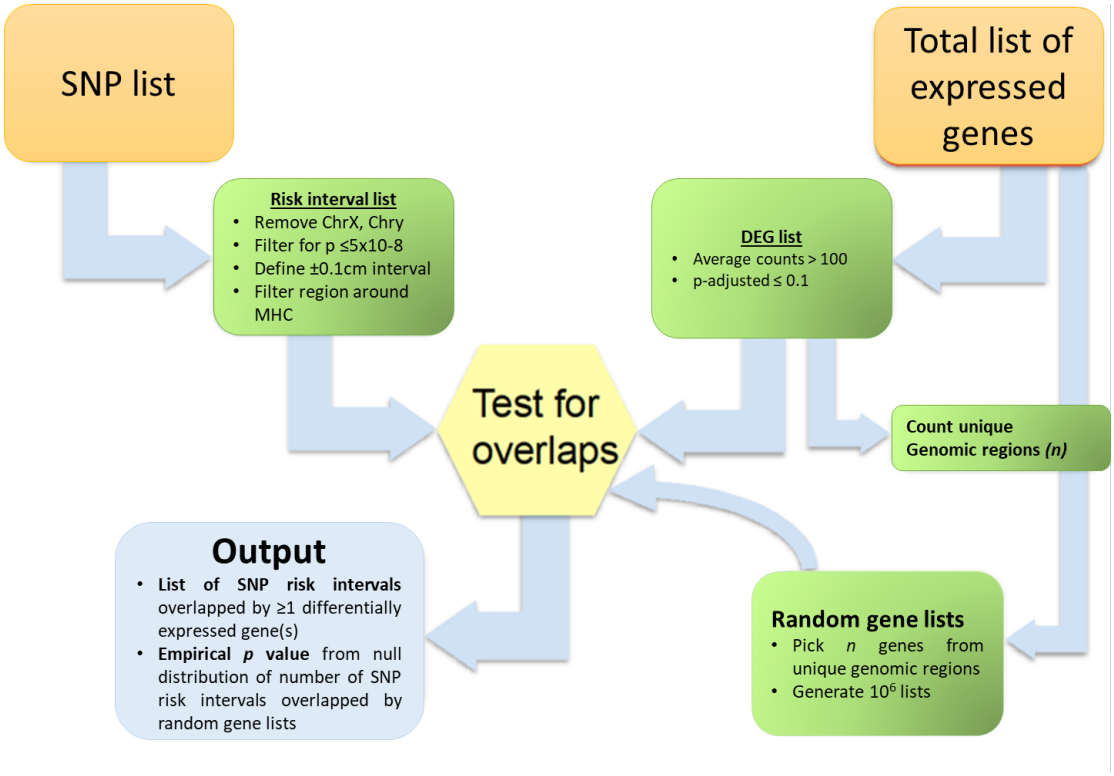


Figure 7.1 OUTLINE OF THE ALGORITHM USED FOR CALCULATING ENRICHMENT IN REGIONS SURROUNDING DISEASE OR TRAIT ASSOCIATED FOCAL SNPS (Raine *et al.*, 2015).

7.3.3 Integration Of UC Risk Associated SNPs With RNA Seq And ATAC Seq Data

First, we took the output of DA peak / SNP enrichment testing and summarized all peaks falling in a pre-defined window around the focal SNP. Next, we identified the peaks most distal to the focal SNP and generated a "dummy" peak with start and end points matching the start and end coordinates for most distal peaks. All peaks, including the "dummy" peak, were annotated; real peaks were annotated to genomic feature and the gene they were closest to, whereas "dummy" peaks were annotated so that all genes in ±1Mb were assigned to the peak. Next, gene lists associated with "dummy" peak were screened for overlap with DEG. Finally, the list of matched hits was filtered to exclude all genes with average >100 counts and peaks with >50 counts. The genomic location of each peak (i.e. if candidate gene is already known) was considered but not used as a filtering criterion. Finally, Pubmed literature searches using

gene name, gene name + cell type, gene name + IBD and gene name + UC were carried out.

7.4 Results

7.4.1 Estimating How Much Of The Chromatin Regulatory Landscape Can Be Inferred From Gene Expression And Vice Versa

Study by *Boyle et al., 2008* showed that highly expressed genes are more likely to have their TSS open (Boyle *et al.,* 2008). As first step in our multi-omics analysis we aimed to evaluate if disease state specific change in expression can be reproduced by the condition specific difference in chromatin conformation and by what extent.

First, we estimated the relationship between the significantly different genes and significantly different peaks. LPL $CD4^+$ T_{EM} showed moderate-to-low, but LPL $CD19^+$ B cells strong association between DEG and DA as indicated by correlation coefficient (*r = 0.3 & r = 0.65*). However, neither of these observed coefficients reached statistical significance (*p = 0.34 & p = 0.06*) (Table 7.1 A). Subsequent, visual assessment of scatter plots revealed that there was no apparent meaningful correlation and these coefficients were indeed likely arising through chance (Figure 7.2 B and C). Blood $CD19^+$ B cells had no relationship between DEG and DA (Figure 7.2 A).

We next evaluated how global difference in chromatin openness (All Peaks) correlates with DEG and how DA regions correlates to global change in gene expression (All Genes). As with the DEG - DA calculations, DEG - All Peaks correlation tests were too low on peak-gene numbers to result in reliable coefficients (Table 7.1 B).

Testing for correlation between All Genes and DA resulted in better numbers of peak-gene pairs. Very low correlation coefficients showed that there was in fact no or minimal relationship between differences in gene expression and DA promoter sites for any of the cell types studied (Table 7.1 C). Interestingly, intronic DA peaks displayed closer relationship with gene expression ($r_{Blood\ CD19+\ B\ cell} = 0.239$; $r_{LPL\ CD19+\ B\ cell} = 0.2292$; $r_{LPL\ CD4+\ TEM} = 0.3144$) than DA peaks assigned to promoter regions ($r_{Blood\ CD19+\ B\ cell} = 0.1119$; $r_{LPL\ CD19+\ B\ cell} = 0.1834$; $r_{LPL\ CD4+\ TEM} = 0.1008$).

Next, we asked if global changes in gene expression are reflected by global variance in chromatin profile. No relationship was seen for peaks within a promoter region, nor any other genomic features tested (Table 7.1 D).

Finally, we decided to repeat the analysis using the previously estimated differences between control blood $CD4^+$ T_{EM} vs $CD19^+$ B cell expression and chromatin profiles. Both control populations had higher sample number and we hoped that it would allow us to capture real cell type specific biological difference.

There was a robust significant correlation between the DEG and DA (*r = 0.57, p = 0.0001)* (Table 7.1 A, Figure 7.2 D). The observed correlation weakened when DEG – All Peaks (*r = 0.3264, p = 0.0001*) were compared. The association fell further when testing for correlation between DA and global gene expression (*r = 0.208, p = 0.0001*) and completely disappeared at the level of global expression vs global accessibility (*r = 0.066, p = 0.0001*). DA peaks within introns showed a moderate correlation with DEG and were stronger than promoters when assessed against global changes in gene expression (Table 7.1 C).

Table 7.1 SPEARMAN CORRELATION BETWEEN *A.* DEG AND DA, *B.* DEG AND GLOBAL CHANGES IN THE CHROMATIN PROFILE, *C.* DA AND GLOBAL CHANGES IN THE EXPRESSION PROFILE AND *D.* GLOBAL CHANGES IN THE CHROMATIN PROFILE AND GLOBAL CHANGES IN THE EXPRESSION PROFILE. *r - Spearman's rho; p - p-value; n - peak-gene pair number; DEG - Significantly different genes; DA - Significantly different accessible regions; All Genes - ± significant genes expressed by the cell type(s) under investigation; All Peaks - ± significant peaks belonging to the cell type(s) under investigation; C – Control; LPL - Lamina propria; T_{EM} – T effector memory; UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon.*

A

	DEG - DA				
	Promoter	**Intron**	**Exon**	**5'UTR**	**Distal**
Blood $CD19^+$ B cells (C vs UC_I)	r = -0.08222 ; p = 0.7700; n = 15;	——	——	——	——
LPL $CD19^+$ B cells (C vs UC_I)	r = 0.65; p = 0.0666; n = 9;	——	——	——	——
LPL $CD4^+$ T_{EM} (C vs UC_I)	r = 0.3007; p = 0.3424; n = 12;	——	——	——	——
Blood ($CD4^+$ T_{EM} vs $CD19^+$ B cells)	r = 0.5715; p = <0.0001; n = 165;	r = 0.386; p = 0.0030; n = 57;	r = 0.01471; p = 0.9607; n = 16;	r = 0.09913; p = 0.5375; n = 41;	r = 0.1745; p = 0.3017; n = 37;

B

	DEG - All Peaks				
	Promoter	**Intron**	**Exon**	**5'UTR**	**Distal**
Blood $CD19^+$ B cells (C vs UC_I)	r = -0.02368; p = 0.8926; n = 35;	——	——	——	r = -0.09119; p = 0.7022; n = 20;
LPL $CD19^+$ B cells (C vs UC_I)	r = 0.3963; p = 0.1157; n = 17;	——	——	——	——
LPL $CD4^+$ T_{EM} (C vs UC_I)	r = 0.1202; p = 0.2029; n = 114;	r = 0.1103; p = 0.3781; n = 66;	——	——	r = 0.2719; p = 0.1036; n = 37;
Blood ($CD4^+$ T_{EM} vs $CD19^+$ B cells)	r = 0.3264; p = <0.0001; n = 363;	r = 0.281; p = 0.0046; n = 100;	r = -0.04407; p = 0.8238; n = 28;	r = -0.02942; p = 0.7725; n = 99;	r = 0.09963; p = 0.4335; n = 64;

C

	All Genes - DA				
	Promoter	**Intron**	**Exon**	**5'UTR**	**Distal**
Blood $CD19^+$ B cells (C vs UC_I)	r = 0.1119; p = <0.0001; n = 2035;	r = 0.239; p = <0.0001; n = 395;	r = 0.0681; p = 0.4796; n = 110;	r = 0.1117; p = 0.0209; n = 427;	r = 0.08471; p = 0.1676; n = 267;
LPL $CD19^+$ B cells (C vs UC_I)	r = 0.1834; p = <0.0001; n = 1871;	r = 0.2292; p = <0.0001; n = 472;	r = 0.2056; p = 0.0209; n = 126;	r = 0.1915; p = 0.0002; n = 386;	r = 0.1242; p = 0.0439; n = 264;
LPL $CD4^+$ T_{EM} (C vs UC_I)	r = 0.1008; p = 0.0498; n = 379;	r = 0.3144; p = 0.0009; n = 108;	——	r = 0.05899; p = 0.5941; n = 84;	r = 0.134; p = 0.2329; n = 84;
Blood ($CD4^+$ T_{EM} vs $CD19^+$ B cells)	r = 0.208; p = <0.0001; n = 1043;	r = 0.2701; p = <0.0001; n = 219;	r = 0.1535; p = 0.1827; n = 77;	r = 0.1795; p = 0.0053; n = 240;	r = 0.1057; p = 0.2137; n = 140;

D

	All Genes - All Peaks					
	Promoter	**Intron**	**Exon**	**5'UTR**	**Distal**	**3'UTR**
Blood CD19$^+$ B cells (C vs UC$_I$)	r = 0.09315; p = <0.0001; n = 4918;	r = 0.08739; p = <0.0001; n = 2345;	r = 0.1357; p = 0.0035; n = 462;	r = 0.06424; p = 0.0300; n = 1141;	r = 0.08471; p = 0.0034; n = 1197;	r = 0.01921; p = 0.8071; n = 164;
LPL CD19$^+$ B cells (C vs UC$_I$)	r = 0.134; p = <0.0001; n = 4150;	r = 0.1593; p = <0.0001; n = 1690;	r = 0.1024; p = 0.0539; n = 355;	r = 0.1167; p = 0.0003; n = 964;	r = 0.07888; p = 0.0164; n = 926;	r = 0.1204; p = 0.1848; n = 123;
LPL CD4$^+$ T$_{EM}$ (C vs UC$_I$)	r = 0.02367; p = 0.0973; n = 4911;	r = 0.1376; p = <0.0001; n = 1250;	r = -0.02832; p = 0.6228; n = 304;	r = 0.0494; p = 0.1105; n = 1045;	r = 0.05019; p = 0.2113; n = 622;	r = 0.2183; p = 0.0399; n = 89;
Blood (CD4$^+$ T$_{EM}$ vs CD19$^+$ B cells)	r = 0.0663; p = <0.0001; n = 4019;	r = 0.1825; p = <0.0001; n = 452;	r = 0.08152; p = 0.2753; n = 181;	r = 0.06894; p = 0.0523; n = 793;	r = 0.03026; p = 0.6016; n = 300;	r = -0.03711; p = 0.8274; n = 37;

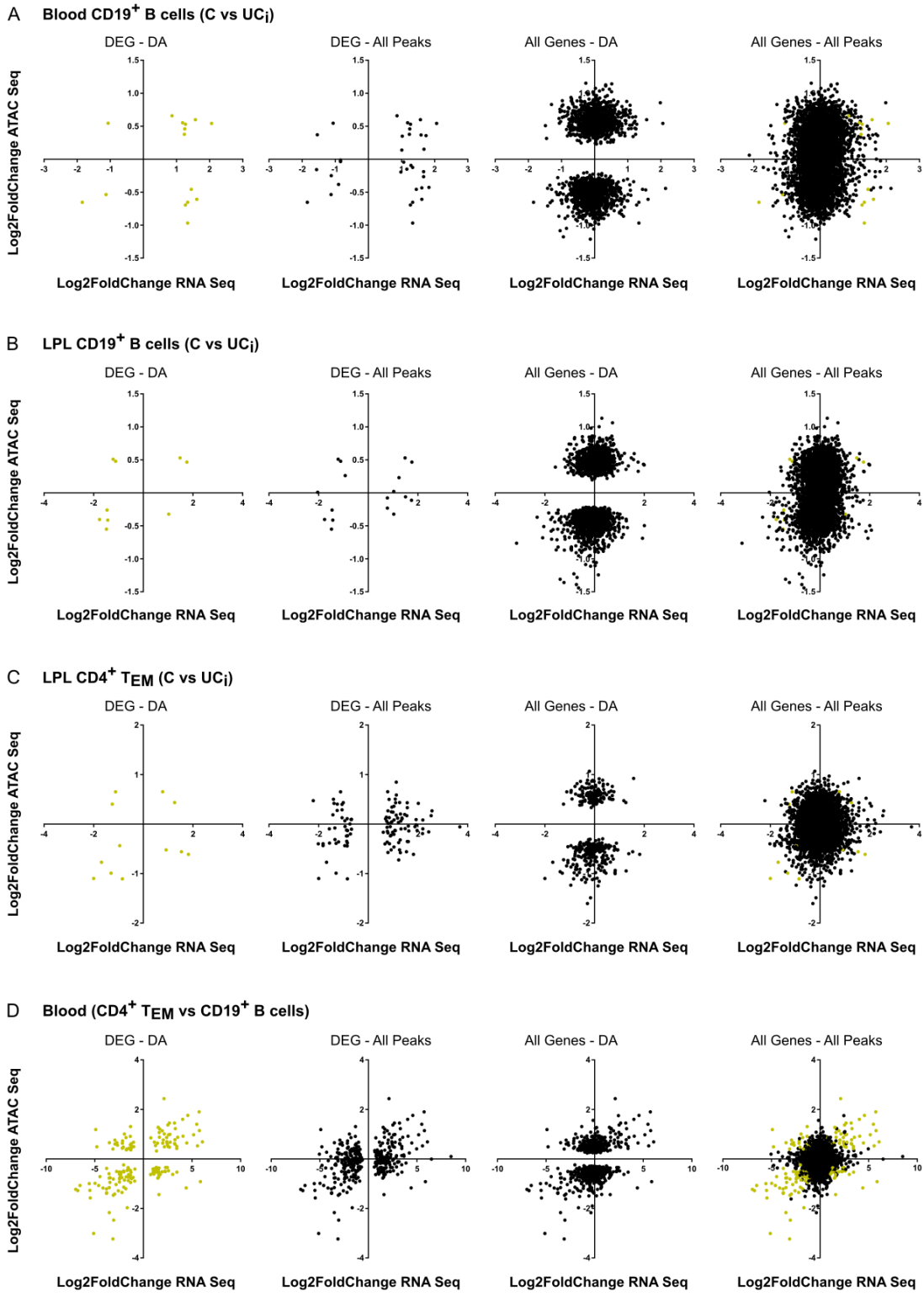


Figure 7.2 VISUAL REPRESENTATION OF RELATIONSHIP BETWEEN GENE EXPRESSION AND PROMOTER ACCESSIBILITY IN *A.* BLOOD CD19^{+} B CELL, *B.* LPL CD19^{+} B CELL AND *C.* LPL CD4 T_{EM} CELLS VARYING BY DISEASE STATE AND *D.* CD4^{+} T_{EM} VS CD19^{+} B CELLS. *The x-axis represents the change in gene expression, whereas y-axis is variation in assigned promoter accessibility. Olive-green dots represent significantly different gene-peak*

LEGEND CONTINUED IN NEXT PAGE

pairs. r - Spearman's rho; p - p-value; n - peak-gene pair number; DEG - Significantly different genes; DA - Significantly different accessible regions; All Genes - ± significant genes expressed by the cell type(s) under investigation; All Peaks - ± significant peaks belonging to the cell type(s) under investigation; C – Control; LPL - Lamina propria; T_{EM} – T effector memory; UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon.

7.4.2 Enrichment Of Disease Associated Genes And Chromatin Regions Within The GWAS Risk Loci Associated With Immune Mediated Diseases Or Traits

We aimed to determine if GWAS identified immune – mediated disease or trait associated loci are enriched for the DEG and DA identified. Scripts used for enrichment estimation were developed by Dr Tim Raine (original method published in Gut 2015) and adapted by author to fit current data sets.

7.4.2.1 Testing For Disease State Specific Expression Enrichment

Immune - disease associated variants, but not trait connected SNPs, approached ($p_{CD} = 0.084$, $p_{RhArt} = 0.06$) or were significantly enriched ($p_{UC} = 0.008$, $p_{Coeliac} = 0.0075$) with genes differentially expressed between LPL CD4$^+$ T_{EM} (C *vs* UC_i). Enrichment strengthened when only genes up-regulated in UC_i were tested (Table 7.2). The majority of genes enriched in UC risk associated regions were either inflammatory (*CCR1, CXCL1, IL-21, IL1R1, RORC, IL2RA, PRDM1, TRAFD1, LY75*) or required for cell proliferation (*NUSAP1, KIFF11, PIM3*).

A large proportion of immune - mediated disease risk loci are shared. Therefore, next we proceeded to determine how much of the enrichment observed between different diseases for genes upregulated in LPL CD4$^+$ T_{EM} UC_i was due to common genetic associations. Association was considered to be shared if immune-mediated disease focal SNPs were in proximity to the same gene and located in the same locus as the UC focal SNP.

As result we identified that 86.67%, 60%, 28.57% and 60% of CD, Coeliac, RhArt and T1D risk loci, enriched for genes up-regulated in active inflammation in LPL CD4$^+$ T_{EM}, were shared with UC. *TRAFD1 (TRAF-Type Zinc Finger Domain Containing 1),* a gene associated with negative control of the immune system (Mashima *et al.*, 2005; Sanada *et al.*, 2008), was identified as overlapping SNPs for all 5 immune-mediated disease. *IL-21*, gene encoding for IL-21 - cytokine predominantly secreted by T cells and naturel killer T cells, was shared between CD, UC, Coeliac and T1D. *CTLA4* and/or *ICOS*

(members of CD28 family of T-cell co-stimulatory receptors) – were up regulated in LPL $CD4^+$ T_{EM} upon active inflammation and associated with risk loci for Coeliac, T1D and RhArt but not UC and CD.

UC associated risk loci might naturally be enriched for immune and inflammatory genes and, thus, the enrichment we observed might simply be due to gene expression data obtained from tissue under conditions of active inflammation. In this regard, DEG showing upregulation in quiescent UC LPL $CD4^+$ T_{EM} cells $_{(C\ vs\ UCn)}$ and approaching significant enrichment for UC associated risk loci ($p = 0.098$) might be more of importance (Table 7.2). We identified a total of 6 genes - *YPEL1, YPEL3, HELZ2, CENPO, SNX27* and *SPREAD2* - enriched in UC risk associated regions. However, *SNX27* was associated with the same locus as *RORC*, already a well-established candidate gene involved as the master transcription factor for pathogenic Th17 cells (Ivanov *et al.*, 2006). Likewise, the locus we assigned to *YPEL3* was already associated with a credible candidate gene ITGAL, identified from eQTL studies in monocytes response to Lipopolysacharide (LPS) stimulus (de Lange *et al.*, 2017a).

Table 7.2 IMMUNE-DISEASE OR TRAIT ASSOCIATED VARIANT ENRICHMENT FOR PEAKS AND GENES IDENTIFIED AS SIGNIFICANTLY DIFFERENT IN CELL POPULATIONS VARYING BY DISEASE STATE OR ANATOMICAL LOCATION. *Enrichment was determined for all DEG/DA and DEG/DA up or down regulated in UC_i/UC_n or SC LPL relative to the Control or Blood. The total number of risk associated regions for each condition, used for enrichment assessment after removal of an overlapping regions, is shown at the head of each column. The number of DEG and DA peak are displayed next to each comparison. p-values for are shown only for observations that were close to (0.1>p>0.05; in blue) or passed the significance threshold (0.05 > p; in red). ALL - all peaks/genes identified as significantly different; UP - DEG/DA upregulated/more open in UC_i/UC_n/LPL; DOWN - DEG/DA downregulated/less open in UC_i/UC_n/LPL; SNP - single nucleotide polymorphism; DEG - Significantly different genes; DA - Significantly different accessible regions; C – Control; UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon; UC(N) - Ulcerative colitis patient with non-inflamed Sigmoid colon; LPL - Lamina propria; T_{EM} – T effector memory; CD - Crohn's disease; UC - Ulcerative colitis; RhArt - Rheumatoid arthritis; T1D - Type 1 diabetes; BMI - Body mass index; MCV - Mean corpus volume.*

				CD	UC	Coeliac	RhArt	T1D	MCV	BMI	Height
RNA Seq		**DEG**	**SNP**	195	180	31	53	28	237	73	158
LPL $CD4^+$ T_{EM} (C vs UC_i)	ALL	288		0.084	0.008	0.0075	0.06				
	DOWN	119									
	UP	169		0.014	0.002	0.0033	0.015	0.055			
LPL $CD4^+$ T_{EM} (C vs UC_n)	ALL	62			0.098						
	DOWN	45									
	UP	17									
LPL $CD19^+$ B cells (C vs UC_i)	ALL	65									0.072
	DOWN	37									0.094
	UP	28									
Blood $CD19^+$ B cells (C vs UC_i)	ALL	69									
	DOWN	22									
	UP	47									0.043
$CD19^+$ B cells (Blood vs. LPL)	ALL	839		0.003	0.003						
	DOWN	180		0.028	0.033		0.099				
	UP	659		0.07	0.049						
$CD4^+$ T_{EM} (Blood vs. LPL)	ALL	520			0.08	0.03	0.031				
	DOWN	178					0.094				
	UP	342			0.035	0.0608					
ATAC Seq		**DA**									
Blood $CD4^+$ T_{EM} (C vs UC_i)	ALL	359									
	DOWN	211									
	UP	148							0.032	0.053	
Blood $CD8^+$ T_{EM} (C vs UC_i)	ALL	2126									
	DOWN	1126									
	UP	1000									
Blood $CD19^+$ B cells (C vs UC_i)	ALL	6008									
	DOWN	3078									
	UP	2930									
IEL $CD8^+$ T_{EM} (C vs UC_i)	ALL	444									
	DOWN	357									
	UP	87			0.035						
LPL $CD4^+$ T_{EM} (C vs UC_i)	ALL	1199									
	DOWN	756									
	UP	443							0.006		
LPL $CD19^+$ B cells (C vs UC_i)	ALL	6551									
	DOWN	3469									
	UP	3082									

7.4.2.2 Testing For Disease State Specific Chromatin Conformation Enrichment

With an exception of IEL CD8$^+$ $T_{EM,}$ none of the immune-mediated disease associated locus showed any enrichment (Table 7.2). Peaks, more accessible in IEL CD8$^+$ T_{EM} in inflamed SC of UC patients showed significant (*p = 0.035*) enrichment for UC risk loci with 8 of 180 focal SNPs having an associated DA peak.

Genetic loci associated with changes in MCV reached significance for overlap with peaks more open in Blood CD4$^+$ T_{EM} and LPL CD4$^+$ T_{EM} in UC under active inflammation.

7.4.2.3 Testing For Cell Linage Specific Expression Enrichment

It has been proposed that genes that are specific to intestinal lymphocytes in comparison to their equivalent populations resident in other locations such as peripheral blood, could be of importance in their identity and function as intestinal lymphocytes, and variation in their regulatory regions could lead/contribute to intestine specific immune disease risk.

DEG genes between control CD4$^+$ $T_{EM\ (Blood\ vs\ LPL)}$ were enriched in UC, CD and RhArt associated risk loci (Table 7.2). Further analysis showed that that genes upregulated in intestinal CD4$^+$ T_{EM} had a stronger association (*p = 0.035*) with UC, whereas coeliac enrichment weakened (*p = 0.06*) than for the full DEG list (that includes both up- and down-regulated genes). Interestingly, it was genes more expressed in Blood CD4$^+$ T_{EM} that approached significant enrichment in RhArt.

In contrast to CD4$^+$ T_{EM} cells, both genes significantly upregulated and downregulated in intestinal B cells were enriched in risk loci associated with UC and CD, but no other immune-mediated disease or trait. Further analysis showed that most of UC risk associated regions enriched for genes upregulated in either blood or LPL residing B cells were shared with CD (92.86% and 78.79%, respectively) (Table 7.3).

Table 7.3 LIST OF GENES THAT WERE IDENTIFIED AS DIFFERENTIALLY EXPRESSED IN B CELLS FROM PERIPHERAL BLOOD COMPARED TO THEIR GUT COUNTERPARTS AND ENRICHED IN BOTH UC AND CD RISK ASSOCIATED LOCUS. *CD - Crohn's disease; UC - Ulcerative colitis.*

CD			UC		
SNP	**CHR**	**SYMBOL**	**SNP**	**CHR**	**SYMBOL**
rs59043219	1	TRAF3IP3	rs59043219	1	TRAF3IP3
rs7555082	1	PTPRC	rs7555082	1	PTPRC
rs4656958	1	SLAMF7	rs4656958	1	SLAMF7
rs12103	1	SDF4	rs12103	1	SDF4
rs3766606	1	ERRFI1	rs3766606	1	ERRFI1
rs4656958	1	LY9	rs4656958	1	LY9
rs59043219	1	G0S2	rs59043219	1	G0S2
rs3024505	1	DYRK3	rs3024505	1	DYRK3
rs34856868	1	GFI1	rs34856868	1	GFI1
rs3024505	1	MAPKAPK2	rs3024505	1	MAPKAPK2
rs7554511	1	INAVA	rs7554511	1	INAVA
rs4656958	1	ITLN1	rs4656958	1	ITLN1
rs12103	1	TNFRSF4	rs12103	1	TNFRSF4
rs925255	2	FOSL2	rs925255	2	FOSL2
rs2593855	3	FOXP1	rs2593855	3	FOXP1
rs2581828	3	SPCS1	rs9847710	3	SPCS1
rs503734	3	NFKBIZ	rs503734	3	NFKBIZ
rs3197999	3	GMPPB	rs3197999	3	GMPPB
rs13126505	4	BANK1	rs13126505	4	BANK1
rs2472649	4	PF4	rs2472649	4	PF4
rs2472649	4	CXCL2	rs2472649	4	CXCL2
rs1363907	5	LIX1	rs1363907	5	LIX1
rs4703855	5	MAP1B	rs4703855	5	MAP1B
rs4976646	5	LMAN2	rs4976646	5	LMAN2
rs1819333	6	CCR6	rs1819333	6	CCR6
rs7746082	6	PRDM1	rs7746082	6	PRDM1
rs7773324	6	IRF4	rs7773324	6	IRF4
rs1456896	7	IKZF1	rs1456896	7	IKZF1
rs1734907	7	ACHE	rs1734907	7	ACHE
rs11768365	7	KDELR2	rs11768365	7	KDELR2
rs7911264	10	HHEX	rs7911264	10	HHEX
rs111456533	10	FAM53B	rs111456533	10	FAM53B
rs111456533	10	AC068896.1	rs111456533	10	AC068896.1
rs11221332	11	ETS1	rs11221332	11	ETS1
rs559928	11	FKBP2	rs559928	11	FKBP2
rs148319899	12	LRRK2	rs148319899	12	LRRK2
rs1569328	14	FOS	rs1569328	14	FOS
rs7404095	16	PRKCB	rs7404095	16	PRKCB
rs28449958	16	NUPR1	rs28449958	16	NUPR1
rs28449958	16	SULT1A2	rs28449958	16	SULT1A2
rs12942547	17	PLEKHH3	rs12942547	17	PLEKHH3
rs12942547	17	RAMP2	rs12942547	17	RAMP2
rs12946510	17	PPP1R1B	rs12946510	17	PPP1R1B
rs6088765	20	SPAG4	rs6088765	20	SPAG4
rs6088765	20	TP53INP2	rs6088765	20	TP53INP2
rs6088765	20	EDEM2	rs6088765	20	EDEM2
rs4256018	20	FERMT1	rs4256018	20	FERMT1
rs6062496	20	ZBTB46	rs6062496	20	ZBTB46
rs2266959	22	SDF2L1	rs2266959	22	SDF2L1

7.4.3 Integrating GWAS, ATAC Seq And RNA Seq Data To Predict Molecular Mechanisms By Which UC Risk Associated Variants Might Contributes To The Studied Phenotype

As a final step in our analysis we tried to combine all 3 - *omics* data sets to narrow down the possible candidate gene list and propose mechanisms by which variants located in the UC associated risk loci might lead to observed changes in expression. We hypothesized that a causal variant located in the risk associated locus might change the function of a regulatory region, such as an enhancer, insulator or silencer, which in turn might lead to change in gene expression. We based our analysis on assumption that risk variants act on chromatin within their immediate proximity. Indeed, the GTEx consortium has showed that in tissue most of eQTLs are in less than 1Mb from gene which expression they affect (GTEx Consortium *et al.*, 2017).

In Blood $CD19^+$ B cells, LPL $CD4^+$ T_{EM} and LPL $CD19^+$ B cells $_{(C\ vs\ UCi)}$ we identified 9, 13 and 12 UC associated risk regions which contained both - DA region and were near to DEG (Table 7.4).

Table 7.4 LIST OF DEG AND DA WHICH EITHER FELL INTO OR WERE IN PROXIMITY TO UC ASSOCIATED RISK LOCUS. *DA peaks and DEG that fall into the same region and share the same direction (in terms of Log2FC) are coloured green (more open and expressed in UC_i) and red (less open and expressed in UC_i).*

The first 2 columns show the GWAS data:

***Column 1** shows focal SNP*

***Column 2** shows chromosome focal SNP is located on;*

Next 9 columns summarize the information around the DA peaks in proximity to focal SNP:

***Column 3** shows DA peak count in each newly defined region 0.1cM each side of focal SNP;*

***Column 4** shows the peak-associated region overlap type;*

***Columns 5** and **6** shows the start and end coordinated of each DA peak;*

***Column 7** shows genomic feature (GF) and gene (in brackets) which peak is associated with based on its location on DNA;*

***Column 8** shows peak width in bp;*

***Column 9**, **10** and **11** shows average concentration (AvgConc), log2foldchange (Log2FC) and p-value adjusted for multiple testing*

Last 4 columns present DEG associated statistics:

***Columns 12**, **13**, **14** and **15** shows average concentration (AvgConc), log2foldchange (Log2FC), p-value adjusted for multiple testing and symbol of differentially expressed gene.*

GWAS - Genome wide association study; ATAC Seq - Assay for Transposase accessible chromatin; RNA Seq - Ribonucleic acid sequencing; SNP - Single nucleotide polymorphism; C - Control, UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon; T_{EM} – T effector memory; LPL - Lamina propria; CHR - Chromosome; No - Number in peaks in region around focal SNP; GF - Genome feature; AvgConc - Average concentration; Log2FC - Log2foldchange; bp - Base pair; LP - little peaks that start and end coordinates of peaks are inside the region around focal SNP;

Blood $CD19^+$ B cells (C vs UC_i)

GWAS		ATAC seq									RNA seq			
SNP	CHR	No	Overlap	Start	End	GF	Width	AvgConc	Log2FC	p-adjust	AvgConc	Log2FC	p-adjust	Symbol
rs75900472	chr9	1	LP	4984368	4985538	Promoter (JAK2)	1171	189.38	-0.39	5.78E-02	8.18	4.35	8.98E-02	INSL6
rs1847472	chr6	1	LP	90304262	90304674	Distal (BACH2)	413	40.86	-0.67	9.50E-02	2618.07	1.16	9.49E-03	UBE2J1
rs1363907	chr5	1	LP	96936722	96936922	Exon (LNPEP)	201	49.31	-0.61	2.86E-02	645.33	1.44	9.49E-03	ELL2
rs2488397	chr1	1	LP	197774700	197775632	Promoter (DENND1B)	933	249.79	-0.70	1.19E-02	338.24	1.27	5.74E-02	DENND1B
		2	LP	197902379	197903037	Promoter (C1orf53)	659	118.28	-0.57	2.52E-02				
rs7555082	chr1	1	LP	198620842	198621819	Distal (MIR181B1)	978	216.60	-0.35	2.44E-02	338.24	1.27	5.74E-02	DENND1B
		2	LP	198622653	198622888	Distal (MIR181B1)	236	57.98	-0.39	8.59E-02				
rs5763767	chr22	1	LP	29766716	29767411	Promoter (UQCR10)	696	79.79	**0.56**	2.92E-02	2221.04	**2.06**	9.49E-03	XBP1
rs2413583	chr22	1	LP	39318999	39319884	Promoter (SNORD43)	886	77.77	**0.62**	7.00E-02	16.55	**3.52**	7.57E-02	DMC1
rs12942547	chr17	1	LP	42275898	42276625	Promoter (STAT5B)	728	37.13	0.66	4.34E-02	210.13	-1.54	9.67E-02	JUP
		2	LP	42288220	42289335	Promoter (STAT5A)	1116	54.99	0.49	6.61E-02				
		3	LP	42387838	42389167	Promoter (STAT3)	1330	233.78	0.45	1.98E-02				
		4	LP	42458520	42458932	Promoter (ATP6V0A1)	413	42.22	0.61	4.85E-02				
		5	LP	42535884	42536687	Promoter (NAGLU)	804	79.38	0.65	1.81E-02				
		6	LP	42566769	42567531	Promoter (MLX)	763	52.95	0.78	3.28E-02				
		7	LP	42608914	42609726	Promoter (TUBG1)	813	113.93	0.78	7.24E-03				
		8	LP	42676960	42677305	Promoter (PLEKHH3)	346	58.73	0.62	3.80E-02				
rs7495132	chr15	1	LP	90529435	90530659	Promoter (CRTC3)	1225	69.56	**0.43**	9.82E-02	409.54	**1.18**	2.70E-02	IDH2
		2	LP	90665614	90665970	Intron (CRTC3-AS1)	357	47.91	**0.72**	3.31E-02				

LPL CD4$^+$ T_{EM} (C *vs* UC_i)

GWAS		ATAC seq									RNA seq			
SNP	CHR	No	Overlap	Start	End	GF	Width	AvgConc	Log2FC	p-adjust	AvgConc	Log2FC	p-adjust	Symbol
rs111456533	chr10	1	LP	124718099	124718501	Intron (FAM53B)	403	56.06	**0.73**	5.63E-02	732.07	**1.23**	4.41E-04	FAM53B
rs7911264	chr10	1	LP	92592870	92593137	Promoter (KIF11)	268	64.96	-0.60	9.85E-02	80.17	2.68	6.87E-02	KIF11
rs653178	chr12	1	LP	111444888	111445184	Intron (SH2B3)	297	27.68	**0.91**	7.13E-02	212.58	**1.29**	4.57E-02	TRAFD1
rs11168249	chr12	1	LP	47832267	47833414	Promoter (HDAC7)	1148	38.77	0.77	7.05E-02	32.00	-4.95	2.53E-02	SLC38A4
rs13001325	chr2	1	LP	102367294	102367856	Exon (IL18R1)	563	57.77	-0.68	6.98E-02	310.46	1.44	7.19E-03	IL1R1
rs1517352	chr2	1	LP	191020137	191020636	Promoter (STAT1)	500	127.40	-0.50	4.52E-02	798.96	0.92	1.77E-02	NAB1
rs10495903	chr2	1	LP	43637073	43637481	Promoter (PLEKHH2)	409	32.59	**-0.70**	9.71E-02	73.62	**-2.40**	9.40E-02	PLEKHH2
rs6142618	chr20	1	LP	32357770	32358639	Promoter (ASXL1)	870	129.27	**0.40**	7.08E-02	562.43	-1.13	8.77E-02	NOL4L
											85.27	**2.73**	1.82E-02	TPX2
rs3197999	chr3	1	LP	49007050	49007605	Promoter (WDR6)	556	114.60	**0.44**	7.74E-02	19.43	**4.69**	7.86E-04	CDC25A
rs1479918	chr4	1	LP	122540006	122540379	Distal (IL21-AS1)	374	31.13	-1.29	2.26E-02	18.37	4.53	5.01E-02	IL21
		2	LP	122578277	122578820	Distal (IL21-AS1)	544	98.33	-0.97	1.12E-02				
rs6863411	chr5	1	LP	142223535	142224014	Distal (SPRY4-IT1)	480	76.00	**-0.62**	9.63E-02	14.55	**-4.35**	7.60E-02	PCDHGB3
rs1847472	chr6	1	LP	90218452	90218949	Intron (BACH2)	498	24.40	**-0.91**	5.46E-02	320.72	**-0.84**	9.49E-02	CASP8AP2
		2	LP	90233861	90234354	Intron (BACH2)	494	53.97	**-0.96**	5.79E-03				
rs10486483	chr7	1	LP	26864310	26864913	5' (SKAP2)	604	144.19	-0.67	4.81E-02	484.36	0.77	3.07E-02	CBX3

LPL CD19$^+$ B cells (C vs UC_i)

GWAS		ATAC seq									RNA seq			
SNP	CHR	No	Overlap	Start	End	GF	Width	AvgConc	Log2FC	p-adjust	AvgConc	Log2FC	p-adjust	Symbol
rs1801274	chr1	1	LP	161539851	161540744	Intron (FCGR2A)	894	46.65	-0.43	9.46E-02	235.464	1.28128	0.01818	FCGR2B
rs4246215	chr11	1	LP	61792011	61792982	Promoter (FEN1)	972	47.64	**0.43**	9.55E-02	35.4793	**3.41917**	0.06842	LBHD1
		2	LP	61815117	61815453	Promoter (MIR1908)	337	23.21	**0.93**	4.89E-03				
		3	LP	61816802	61817159	5' (FEN1)	358	29.51	**0.66**	1.54E-02				
		4	LP	61891067	61892246	Promoter (FADS3)	1180	140.78	**0.30**	7.01E-02				
rs28449958	chr16	1	LP	28292077	28292628	Promoter (SBK1)	552	42.24	0.67	1.54E-02	67.9471	-2.6589	0.04357	SULT1A1
		2	LP	28507658	28508345	Intron (MIR3680-1)	688	35.13	0.60	1.87E-02				
		3	LP	28553686	28554401	Promoter (SGF29)	716	37.15	0.68	6.53E-03				
		4	LP	28822517	28823117	Promoter (ATXN2L)	601	45.66	0.52	4.54E-02				
		5	LP	28823637	28824602	5' (MIR3680-1)	966	94.06	0.75	1.83E-03				
		6	LP	28846083	28846699	Promoter (SH2B1)	617	108.77	0.55	1.70E-02				
		7	LP	28863164	28863867	5' (MIR3680-1)	704	72.77	0.91	4.28E-05				
		8	LP	28879664	28880247	Promoter (ATP2A1-AS1)	584	38.97	0.53	4.50E-02				
		9	LP	28923832	28925811	5' (MIR3680-1)	1980	196.03	0.42	4.40E-02				
		10	LP	28945599	28945803	Distal (MIR3680-1)	205	19.91	0.66	4.65E-02				
		11	LP	28950480	28951065	Promoter (NFATC2IP)	586	113.89	0.47	2.80E-02				
rs1728785	chr16	1	LP	68448184	68448825	Promoter (SMPD3)	642	117.72	0.28	9.37E-02	56.4635	-2.9688	0.00686	NIP7
		2	LP	68470028	68470852	Distal (SMPD3)	825	63.35	0.61	2.43E-02	17.5377	-4.7418	0.08938	DPEP3
		3	LP	68538867	68539559	5' (ZFP90)	693	35.83	0.63	8.18E-03				
rs12942547	chr17	1	LP	42388246	42389167	Promoter (STAT3)	922	160.71	0.53	3.98E-04	80.1934	-2.3509	0.01451	PLEKHH3
		2	LP	42535883	42536795	Promoter (NAGLU)	913	87.39	0.63	2.15E-03				
		3	LP	42566852	42567646	Promoter (MLX)	795	41.34	0.43	7.54E-02				
		4	LP	42577600	42578109	Promoter (PSMC3IP)	510	46.10	0.56	4.72E-02				
		5	LP	42578367	42578896	Promoter (PSMC3IP)	530	57.21	0.76	1.56E-02				
		6	LP	42676928	42677303	Promoter (PLEKHH3)	376	80.74	0.54	1.03E-02				
		7	LP	42744372	42745163	Promoter (EZH1)	792	61.02	0.73	6.96E-04				
rs4802307	chr19	1	LP	46346643	46347339	Promoter (PPP5C)	697	35.69	0.67	1.51E-02	11698	-0.9281	0.08792	FOSB
rs144344067	chr2	1	LP	186485769	186486766	Promoter (ZC3H15)	998	248.66	**-0.42**	1.00E-02	12.8012	**-5.7267**	0.06974	FAM171B
		2	LP	186589677	186590336	Promoter (ITGAV)	660	168.82	**-0.35**	7.45E-02				
		3	LP	186693786	186694384	Promoter (FAM171B)	599	31.04	**-0.79**	2.02E-02				
rs1479918	chr4	1	LP	122151979	122152614	Promoter (KIAA1109)	636	162.27	-0.50	5.08E-03	36.8154	4.00369	0.01849	TNIP3
rs4976646	chr5	1	LP	177311538	177312513	Promoter (RAB24)	976	79.46	**0.61**	9.97E-03	18.2929	**5.62202**	0.01818	HK3
		2	LP	177357708	177358227	Promoter (RGS14)	520	40.24	**0.59**	2.62E-02				
		3	LP	177412526	177413108	Intron (GRK6)	583	22.70	**1.01**	3.50E-03				
rs4957048	chr5	1	LP	472488	473378	Promoter (SLC9A3-AS1)	891	66.77	**0.63**	2.82E-03	91.3756	**2.82642**	0.01849	CCDC127
		2	LP	611732	612519	Promoter (LOC100996325)	788	52.16	**0.55**	1.36E-02				
rs10065637	chr5	1	LP	56142411	56143441	Intron (ANKRD55)	1031	229.62	**-0.65**	7.07E-03	275.203	**-1.8522**	0.0176	MIER3
		2	LP	56148296	56148870	Intron (ANKRD55)	575	63.23	**-0.77**	1.44E-02				
		3	LP	56156670	56156958	Exon (ANKRD55)	289	39.27	**-0.91**	1.93E-02				
rs2538470	chr7	1	LP	148501448	148502068	Distal (C7orf33)	621	41.77	**-0.56**	9.59E-02	41.654	**-3.2419**	0.08792	ZNF786

7.5 Discussion

We herein integrated transcriptomics, GWAS and chromatin conformation data for possibly IBD implicated cell types purified from Blood and SC biopsies from healthy controls and UC patients with and without inflammation. First, we would like to acknowledge that RNA seq and ATAC seq data sets are suboptimal quality. Therefore, data in this chapter (on their own) are not reliable to make any biological conclusions. Instead, we proceeded with analysis and interpretation to give author a chance to increase skill set.

7.5.1 Relationship Between Expression And Accessibility

First, we looked if difference in expression profile between C *vs* UC_i are reflected by difference in peak accessibility. We used peak-gene pair model and performed a series of association calculations between DEG – DA, DEG – All Peaks, All Genes – DA and All Genes – All Peaks. Unfortunately, due to low peak-gene pair numbers we were not able to properly assess if and how much significantly different changes in gene expression correlates with significantly different changes in their promoter (or other genomic region) accessibility in the same cell type varying only by disease state.

By using cell type specific data from cell populations of different lineages from healthy controls, we showed that in this context DEG show the strongest correlation to DA regions mapped to promoters. We also demonstrated that DEG exhibit a stronger association with differential promoter site openness than the converse (i.e. than DA promoters being associated with genes showing differential expression). Whether this also applies to DEG and DA regions for the same cell type under different inflammatory states remains unclear. The lack of correlation we observed may reflect the fact that regulation of gene expression under conditions of inflammation is mainly driven by other factors, such as transcription factor occupancy of already open regions of chromatin.

The relationship between expression and chromatin profile are complex. Even though we modelled with the assumption that individual peaks might be responsible for

changes in gene expression and achieved, in the context of an unknown, non-linear biological system, strong correlation, it is not always the case. *de la Torre-Ubieta et al., 2018* looked at difference in chromatin conformation and gene expression between cortical plate and germinal zone of developing human neocortex. They reported only moderate correlation between DA promoters and genes with significantly different expressed exons (*r = 0.417, p = 6.2e-60*).

Correlations between promoter accessibility and gene expression are perhaps the best understood model. In this system, TF binding initiates nucleosome displacement for transcription machinery to bind (Workman and Kingston, 1992; Svaren *et al.*, 1994). Even in this system the relative importance of TF binding within the promoter region can vary depending on the biology under investigation. The impact and mechanism of other regulatory regions on gene expression is less clear. In particular, recent studies by *Alasoo et al., 2018* have shown that a large proportion of regulatory regions can be primed under normal conditions, meaning that allele specificity will impact the chromatin accessibility under normal conditions. This observation means that regions important for disease risk might not be identified from case control studies. In addition, *Alasoo et al., 2018* revealed that some peaks act as master regulators for others. *Gate et al., 2018* showed that a single variant can affect nearby and distal chromatin conformation which might partially explain why one peak can impact the accessibility of other peaks. We believe with increase in information, more complex models will be developed which will permit the identification of the global relations between the gene expression and chromatin conformation.

7.5.2 Disease And Trait Associated Locus Enrichment For DEG And DA Regions

Next, we evaluated if GWAS risk variants associated with immune – mediated diseases or traits are enriched in genes/peaks identified as significantly different between C *vs* UC ($_i$ or $_n$), or Blood *vs* LPL.

We found that 4 out of 5 autoimmune disease associated risk variants, but not trait associated variants, were significantly enriched for genes up-regulated in LPL $CD4^+$ T_{EM}

upon active inflammation, whereas the 5th narrowly failed to reach statistical significance ($p_{T1D} = 0.055$). Further assessment showed that *TRAFD1,* a gene with expression upregulated in LPL CD4$^+$ T_{EM} UC_i, was shared between all 5 autoimmune conditions.

TRAFD1, also known as *FNL29,* was first identified as a IFN-γ + LPS inducible gene in monocytic M1 cells; mouse peritoneal macrophage and macrophage-like RAW cell stimulation with LPS, IFN-β and IFN-γ led to increased TRAFD1 protein levels. Further experiments showed that TRAFD1 overexpression attenuated NF_kB signalling which most likely is the consequence of TRAFD1 physical interaction with TRAF6 (an adaptor protein with multiple functions, from which one is mediating the downstream signal of pattern recognition receptors, such as, toll-like receptors) (Mashima *et al.*, 2005). *Sanada et al., 2008* showed that *Trafd1* deficient mice, challenged with sublethal dose of LPS or Poly(I:C) (Polyinosinic – polycytidylic acid; synthetic analogue of double-stranded RNA, used to study anti-viral pattern recognition receptor signaling) exhibits much higher mortality than WT *in vivo.* In addition, Bone marrow-derived dendritic cells (BMDCs) from *Trafd1* -/- mice secreted significantly higher levels of pro-inflammatory cytokines such as TNF-α, IL-6 and IL-12p70 (an active heterodimer of IL-12) upon stimulation with LPS than BMDCs from WT mice (Sanada *et al.*, 2008). Together they have proposed TRAFD1 as negative regulator of immune response. *TRAFD1* was significantly upregulated in LPL CD4$^+$ T_{EM} in UC patients with active inflammation, and even though physiological role of *TRAFD1* in T cells is still unknown, T cell specific deletion of its proposed interaction partner - TRAF6 results in multiorgan inflammatory response (King *et al.*, 2006), suggesting that TRAFD1 might act as negative feedback mechanism to regulate active inflammation.

In contrast to differential expression data, none of immune-mediated disease showed enrichment for regions displaying higher accessibility in LPL CD4$^+$ T_{EM} UC_i. Which might contradict other already published studies. *Farh et al., 2015* showed that their identified, UC associated variants are enriched for acetylated cis-regulatory elements (which are markers of active promoters and enhancers) on stimulated CD4$^+$ T cells and colonic mucosa. It was seconded by *Kundaje et al.*, 2015b. In addition, *Gate et al., 2018*

showed that both - naïve and stimulated $CD4^+$ T cell specific peaks, show higher overlap with UC associated variants than predicated by chance.

Lack off enrichment possibly could be attributed to poor quality of our ATAC seq data. However, we could not exclude probability that our enrichment model was not fit for ATCA seq data analysis, mainly as individual gene can have multiple peaks with only some reaching statistical significance. Refining our enrichment algorithm might allow us to pick up any missing relationship, yet, we recognized that not all risk associated loci will be effective in all cell types.

An additional approach to identify tissue or cell specific gene expression effects that could have a potential role in pathogenesis involves comparing the transcriptional profiles between matched cell populations (by surface marker expression) in different parts of the body or between case and control samples. The underlying hypothesis is that genes showing differential expression may be of particular importance for the tissue or cell type under study, and that, where these genes lie near to focal SNPs identified by GWAS, they represent attractive candidate genes for further study within that particular locus/cell type combination (Gautier *et al.*, 2012).

Herein we showed that genes upregulated in LPL $CD4^+$ T cells (when compared to their blood counterparts) are enriched UC risk variants. In this regard, the observation that genes upregulated in SC $CD4^+$ T_{EM} but not TI $CD4^+$ T_{EM} (Raine *et al.,* 2015) (in comparison to Blood $CD4^+$ T_{EM}) were strongly enriched in UC associated risk loci suggests a degree of anatomical specificity that matches the known disease distribution. We therefore determined how many genes, which were upregulated in SC $CD4^+$ T_{EM} cells compared to blood and located in proximity to UC associated focal SNPs, were found differentially expressed between C vs UC ($_i$ or $_n$). We found that 5 genes (*PLEKHH2, OTUD3, FOS, TPPP* and *SPRED2*) constituting 20% of genes examined, were identified as differentially expressed between the healthy and disease. Interestingly, all 5 were downregulated in UC.

7.5.3 Integration Of Functional Genomics Data With GWAS Identified UC Associated Risk Variants

Finally, we hypothesized that UC risk associated loci affects regulatory region function which subsequently leads to changes in gene expression. *Gate et al., 2018* assessed chromatin structure in CD3 and CD28 stimulated $CD4^+$ T cells from 105 healthy individuals with European descent. First, they showed that only 5% of SNP-containing peaks had associate SNP affecting chromatin state. However, these peaks were enriched for T cell enhancers and T cell development and activation associated TFBs. Moreover, peaks containing chromatin quantitative treat locus were more likely to overlap with GWAS autoimmune-disease associated risk variants than general SNP-congaing peaks. 30% of local chromatin quantitative treat loci acted as eQTL to nearby genes.

Here we combined ATAC Seq, GWAS and RNA Seq data to propose the causal mechanistic by which variants contained in risk loci might impact the nearby (± 1Mb) gene expression with respect to specific cell types, namely Blood $CD19^+$ B cells, LPL resident $CD19^+$ B cells and $CD4^+$ T_{EM} cells. After screening for DA peaks and DEG in proximity to focal SNPs associated with UC we further filtered our lists to exclude genes and peaks associated with low counts and known association. We were left with 1 overlap region for LPL $CD19^+$ B cells, 4 for Blood $CD19^+$ B cells and 5 for LPL $CD4^+$ $T_{EM.}$ In the following text, each predicated overlap region is discussed separately.

Blood $CD19^+$ B cells

MIER3 (MIER Family Member 3)

We found 3 peaks in region around rs10065637, all showing significantly less accessibility in UC_i. All peaks were located on gene body of ANKRD55. *MIER3* expression was downregulated in UC_i, letting us hypothesize that either of peaks could be harbouring an enhancer site. PubMed search for MIER3 resulted in total of 6 studies. MIER3 expression has been associated with colorectal cancer: *Pitule et al., 2013* compared biopsies from healthy mucosa vs cancer affected areas from CRC patients and reported significant reduction in MIER3 expression in tumour sites. *Peng et al., 2017* further

revisited MIER3 with aim to better understand its role in CRC. They showed that MIER3 downregulation is associated with more aggressive cancer growth and metastasis. Currently there are no studies of MIER3 in B cells, but profiling the tumor environment at single cell level has captured B cells and confirms MIER3 expression in these cells (Jerby-Arnon *et al.*, 2018).

LPL CD19^{+} B cells

DENND1B (DENN Domain Containing 1B)

Two separate loci contained peaks within 1 Mb of *DENND1B*. Each region contained 2 peaks, from which one was located on the *DENND1B* promoter. Although DENND1B expression was increased in UC_i compared to Control, all peaks showed significant decrease in openness. Of the other 3 peaks, one was within the promoter sequence of the *C1orf53* gene. The other 2 were in an intergenic region distal to *MIR181B1*, which encodes a short non-coding microRNA that acts on mRNA stability. Currently, without any further investigation, it is hard to say if/which of peaks identified might have an effect on *DENND1B* expression, yet the decreases in *DENND1B* promoter accessibility might fit the least well. DENND1B functions as a guanine nucleotide exchange factor for small GTPases, specifically Rab35 which is regulator for endocytic recycling (Yoshimura *et al.*, 2010). *Yang et al., 2016* have recently showed that aerosolized antigen challenged Dennd1b-/- mice intensified T helper 2 mediated inflammation. The DENND1B role in B cells is yet to be eluded. Polymorphisms in DENND1B have been associated with childhood asthma (Sleiman *et al.*, 2010).

XBP1 (X-Box Binding Protein 1**)**

The region around focal SNP rs5763767 contains a DA peak within the *UQCR10* promoter. This sequence was more accessible in UC_i. The peak is also around 1Mb from *XBP1*, the expression of which is also increased expression in UC_i. *UQCR10* encodes a component of the ubiquinol-cytochrome C reductase complex (Schägger *et al.*, 1995). Interestingly, when *Takata et al., 2010* compared gene expression in the brain cortex and hippocampus of *Xbp1* +/− *vs* WT littermates, the top genes upregulated were *Uqcr10* and *Nipsnap1*. However, this study was potentially flawed by possible gene

carry over from the 129S mice strain they used to generate *Xbp1*+/- mice. XBP1 is a transcription factor extensively studied in IBD, particularly CD (Kaser *et al.*, 2008; Adolph *et al.*, 2013). In the epithelium, accumulation of misfolded protein can induce endoplasmic reticulum (ER) stress and subsequently lead to the so-called "unfolded protein response" (UPR) (Chakrabarti, Chen and Varner, 2011). IRE1 is one of the ER stress sensors and splices XBP1 mRNA to induce generation of its active form, which now acts as transactivator for UPR target genes. In B cells XBP1 is required for plasma cell development (Reimold *et al.*, 2001).

JUP (Junction Plakoglobin aka Catenin Gamma)

The region associated with *JUP* contains 8 DA peaks, all on promoters showing increased accessibility, whereas *JUP* expression was significantly reduced in UC_i. JUP belong to the catenin family and is vital for desmosomal assembly. In addition it plays a role in cell adhesion and motility (Lie, Cheng and Mruk, 2011). The JUP impact on Wnt signalling is less straight forward than its relative β – catenin (Miller *et al.*, 2013). Both proteins' roles have been studied in IBD as both interact with E-cadherin, an epithelial adhesion molecule. *Karayiannakis et al., 1998* stained colonic tissue from UC and CD patients and showed no difference in JUP expression in both disease. *Koch et al., 2008* looked at JUP and β – catenin impact on haematopoietic stem cell self-renewal and differentiation into other blood cells and concluded that neither are of importance.

IDH2 (Isocitrate Dehydrogenase (NADP(+)) 2, Mitochondrial)

There were two peaks, one on the *CRTC3* promoter region, another on an intron. Both showed increased accessibility in UC_i and were also close to *IDH2*, a gene upregulated in UC_i, that encodes a protein necessary for conversion of isocitrate to α-ketoglutarate (Jo *et al.*, 2001; Lee *et al.*, 2007). IDH2 mutations have been associated with various cancer types, including B cell malignancies (Yang *et al.*, 2012). *Cha et al., 2017* used DSS to induce acute colitis in IDH2+/+ and IDH2-/- mice and showed significant drop in survival rate and increased neutrophil infiltration within the colon of knock out animals. They proposed that IDH2-/- exacerbated the DSS – induced colitis via alteration

in redox status, that leads to increase in p53 upregulated modulator of apoptosis (PUMA)-mediated apoptosis.

LPL $CD4^+$ T_{EM} cells

FAM53B (Family With Sequence Similarity 53 Member B**)**

The risk loci around rs111456533 contained a DA open peak within the intronic portion of the *FAM53B* gene. In parallel, expression of FAM53B was significantly increased. There is little information about FAM53B in the literature currently, but the few existing papers have suggested that the FAM53B orthologue Smp is important in tissue regeneration and cell proliferation in fish (Kizil *et al.*, 2009). Further experiments from *Kizil et al., 2014* showed β-catenin dependence on Smp activity, where loss of Smp prevented the nuclear accumulation of β-catenin and blocked expression of Wnt target genes. They further used FAM53B siRNA transfected HEK293T cells line and showed reduction in β-catenin nuclear translocation in response to stimulus with Wnt3 conditioned medium. *Ding et al., 2008* reported anti-inflammatory effect of stable β-catenin expression in T cells, particularly T_{regs}. The exact T_{reg} role in UC is not fully understood. However, patients with mutations in FOXP3, a key gene for T_{reg} function, leading to dysfunctional T_{regs} are associated with intestinal inflammation (Okou *et al.*, 2014). Hence, FAM53B might be one of the genes necessary for T_{reg} suppressive and inflammatory function and may be impacted in these cells by the disease associated variant.

NAB1 (NGFI-A Binding Protein 1**)**

NAB1 was significantly upregulated in UC_i. NAB1 acts as transcriptional repressor for EGR1 (Early Growth Response-1) (Thiel *et al.*, 2000) and currently there are no associations between NAB1 and IBD. It is located in proximity to DA promoter of *STAT1* (Signal Transducer and Activator of Transcription 1), less accessible in UC_i. STAT1 acts as a TF for Interferon dependent gene expression (Villarino *et al.*, 2015).

NOL4L (Nucleolar Protein 4 Like**)**

NOL4L and *TPX2* are two genes within the rs6142618 region. Only *NOL4L* passed the counts threshold introduced. We observed significant upregulation in $CD4^+ T_{EM}$ in UC_i. NOL4L function in humans is not well understood. *Guastadisegni et al., 2010* and *Kawamata et al., 2012* have identified NOL4L as a fusion partner to RUNX1 and PAX5 in acute myeloid leukemia and acute lymphoblastic leukemia. NOL4L and TPX2 were also in proximity to DA promoter of *ASXL1*, which was more accessible in UC_i.

CASP8AP2 (Caspase 8 Associated Protein 2**)**

The region around rs1847472 contained two DA peaks that were intronic to *BACH2.* Within this region, lies CASP8AP2. The DA peaks showed decreased accessibility in UC_i and CASP8AP2 was downregulated. CASP8AP2 is a multifunctional protein and has been associated with histone gene expression, cell death and survival. It is involved in Fas and TNF- α death receptor signalling, with one of the functions being activation of caspase 8 (CASP8) (Imai *et al.*, 1999; Jun *et al.*, 2005). In addition, *CASP8AP2* has been identified as crucial for CRC cell survival (Hummon *et al.*, 2012). Apoptotic pathways have been long recognized in IBD, and very recently *Lehle et al., 2019* showed that germline mutation in CASP8 which results in reduced protein abundance, is associated with very early onset development of Inflammatory Bowel Disease.

CBX3 (Chromobox 3 **)**

CBX3 expression was increased in UC patients with active inflammation in comparison to healthy control. The gene is located close to a DA peak that lies within the 5'UTR of the *SKAP2* gene. This peak shows significant reduction in chromatin accessibility in UC patients. CBX3 has been associated with various functions including transcriptional regulation (including inflammatory genes) and can act as component of heterochromatin. Sohn *et al.*, 2012 used CBX3 as marker of senescence in UC and CD. Senescence is a process in which cells enters permanent growth arrest. IHC revealed high level of CBX3 in colonic tissue from IBD patients, where CD had higher levels of CBX3 then UC. The role of CBX3 in lymphocytes is less studied. *Sun et al., 2017* looked at Cbx3 effects

on $CD8^+$ T_E cells. They showed that Cbx3 insufficient $CD8^+$ T cells had increased killing capacity and that the cancer microenvironment of Cbx3 insufficient mice or wild type mice treated with Cbx3 insufficient $CD8^+$ T cells, showed changes in T cell ratios, with decrease of $CD4^+$ T_{reg} and increased $CD8^+$ T_E cells.

8. Discussion

UC is chronic idiopathic condition, characterized by uncontrolled inflammatory response in colon (Gramlich and Petras, 2007; Fatahzadeh, 2009). GWAS has been successful in identification of one of the largest numbers of UC and Crohn's disease risk associated regions across the entire human genome, more than for any other immune mediated disease. However, as for other immune mediated diseases, only a small fraction of variants is located in protein-coding regions (de Lange *et al.*, 2017). Though missense variants have been essential in highlighting some of disease associated causal genes and pathways (Pidasheva *et al.*, 2011; Strober and Watanabe, 2011; Adolph *et al.*, 2013; Lamas *et al.*, 2016), the exact biological role of majority of loci still remains unclear.

Previous attempts to understand the molecular paths leading to complex disease have shown that GWAS identified disease and trait associated variants are significantly enriched in endonuclease accessible chromatin regions (Maurano *et al.*, 2012; Kundaje *et al.*, 2015b), chromatin marks (Ernst *et al.*, 2011; Farh *et al.*, 2015; Kundaje *et al.*, 2015b), eQTLS for complex diseases (Nicolae *et al.*, 2010) and histone-QTLs (Grubert *et al.*, 2015). Taken together, these studies proposed that disease associated risk variants could act by modulating the function of underlining regulatory elements, which in turn could lead to change in gene expression. *Grubert et al., 2015* carried out a large scale study looking at genetic variant impacts on histone marks, DHS, gene expression and chromatin 3-dimensional interactions. They showed that a single QTL can affect multiple molecular phenotypes, and that the biggest proportion of local-QTLs affect the enhancer chromatin state. Enhancers are the regulatory elements characterized by their cell-type and tissue specificity (Heinz *et al.*, 2015). A number of studies have now showed strong GWAS risk associated variant enrichment in enhancer sites (including IBD) (Ernst *et al.*, 2011; Farh *et al.*, 2015). Interestingly, it is cell type and tissue specific enhancers which shows the enrichment. *Weedon et al., 2014* revealed that homozygous mutation in distal enhancer of PTF1A gene, were associated familiar pancreatic cancer. Enhancer region showed marked cell-type specificity, as none of the cell types screened by ENCODE showed the enhancer.

The main goal we set with this study was to develop hypotheses around the exact mechanism by which the GWAS associated variants might lead to disease risk. GWAS on its own is good in identifying common disease susceptibility loci without prior knowledge of locus function, but it is unable to identify the causal variant, causal regulatory region or causal gene. From large numbers of common risk regions, we believe that complex networks of molecules are involved in UC pathogenesis. Each genomic study is limited to one particular molecular layer. Hereby, to gain a complete understanding of biological mechanism of disease multiple genomic studies should be employed. We selected RNA Seq and ATAC Seq methodologies for functional analysis of UC. On their own they can highlight certain biological differences between the healthy participants and UC, but we hoped that integrative approach would provide us with understanding of the molecular information flow. Thus, we hoped that we could connect the change in underlying genomic sequence, to change in chromatin conformation, which would further explain the transcriptional differences. However, as at the time there were no studies looking at the expression profiles and chromatin conformation in purified cell populations from gut of either healthy and UC patients, we first proceeded with individual genomic data analysis.

RNA Seq of $CD4^{+}T_{EM}$ and $CD19^{+}B$ cells from blood and SC allowed us to identify 2042 DEG including 5 long non-coding RNAs, between the control and UC $_{(i \text{ and } n)}$. However, it is important to note that sample QC indicated that there might have been a slight contamination during FACS sorting. In addition, Power Analysis showed that with our sample numbers and sequencing depth we have power to reject the null hypothesis only for genes with high expression and LogFold change. Altogether suggesting that data validation is crucial before any biological conclusion is made. Nevertheless, we successfully picked up already known transcriptional difference, such as increased expression of *RORC, IL17A, IL21, IL2RA* and *CTLA4* in LPL $CD4^{+}T_{EM}$ form UC_i. Moreover, we were able to propose new genes which had no or little association with UC. Good examples are *SEMA4A* and *KSR2*. Moreover, we saw that DEG from LPL $CD4^{+}T_{EM\ (C\ vs\ UCi)}$ are enriched in GWAS risk loci for UC. Interestingly, the DEG from the same cell populations approached significant enrichment when control was compared to UC_n.

Next, we investigated the chromatin profile in 10 FACS-purified cell populations from peripheral blood and SC from healthy and diseased. We identified a total of 393'931 accessible regions present in at least 1 of 30 phenotypes. Unfortunately, even after extensive trouble shooting DA calculation returned data with possible quality issues. Thought the small sample number is very unlikely to represent true estimate of population mean (assumption is based on observation that RNA seq data, which had less variance and higher sample counts could reliably identify only very prominent change in gene expression), change of differential accessibility algorithm could solve the failure in test statistics seen.

As both RNA seq and ATAC seq data sets were of suboptimal quality, we were not able to use them for sophisticated computer modelling of the possible causal network, which was our main goal. We still proceeded with multi -*omics* analysis to improve authors skillset. Integration of all three -*omics* data sets allowed us to identify 10 GWAS risk regions where the DA peak and DEG were in near proximity.

For example, in B cells r7495132 SNP were located in close proximity to *IDH2* and two differentially accessible peaks falling in the *CRTC3* promoter and intronic regions. DEG and DAR showed increased chromatin accessibility and IDH2 expression in UC_i. Interestingly, both genes encode for protein important in the energy metabolism. CRTC3 is a downstream element of second messenger cAMP mediated signalling cascade. CRTC3 acts as a co-activator for cAMP responsive element binding protein (CREB) (Liu *et al.*, 2018). Double knockout of *CRTC3* and *CRTC2* is lethal, whereas *CRTC2*-/-, *CRTC3*-/+ resulted in splenomegaly and abnormal bone marrow functions, which included reduction in B cell numbers in circulation (Kim *et al.*, 2017). In addition, CRTC3 regulates mitochondrial biogenesis in response to rotenone induced mitochondrial stress (Than *et al.*, 2011).

The similar multi -*omics* approaches have been successfully employed in studying the GWAS SNP effects in pancreatic cell subtypes isolated from human donors. *Arda et al., 2018* showed that cell type specific DAR are enriched for diabetes associated genes

and that genes expressed in cell type specific manner can be associated with multiple DAR (Arda *et al.*, 2018).

In conclusion, we aimed to study the expression profiles and chromatin accessibility in purified, possibly UC relevant cell populations, with main goal to use these data to propose the functional role of UC associated risk loci. We have generated valuable data sets, which when combined with other data sets, could provide with reliable biological insights. However, we fell that the main impact of this study was many technical findings regarding sample processing and analysis. Since our early work, there have been several publications highlighting the same challenges we have seen, such as read alignment with ribosomal library preparation and ATAC seq data analysis guidelines. Hereby, our work could be used as guidance of how to design better experiments using clinical samples of suboptimal quality.

Unfortunately, we did not have any time to fully validate any of our findings. If we had more time and funds available we would try to overcome some of the biggest limitations in the study - small sample numbers and low sequencing depth. The best starting point would be *in silico* validation of the expression data. *Smillie et al., 2018* compare the expression profiles of UC vs Control at single cell level and, hereby, this dataset could form a resource of first choice.

As for the next steps, we are not aware for any ATAC or DNase Seq assay in purified populations from human intestinal tissue. Unfortunately, currently the only method to validate the ATAC seq data would be the performance of more ATAC seq, DNase seq or Chip Seq for histone marks.

In addition, by end of project there were some ATAC seq libraries leftovers, therefore with more money available, we could increase sequencing depth to the 300M reads/sample suggested, which would allow to gain better insight of chromatin conformation and perform the TF “foot printing” and identify the TFBS occupancy in these samples, which from our current knowledge currently is not available anywhere else.

9. References

1. Adamczyk, A. *et al.* (2017). 'Differential expression of GPR15 on T cells during ulcerative colitis., *JCI Insight*, 2(8), e90585. doi: 10.1172/jci.insight.90585

2. Adolph, T. E. *et al.* (2013) 'Paneth cells as a site of origin for intestinal inflammation', *Nature*, 503(7475), pp. 272–276. doi: 10.1038/nature12599.

3. Al, G. et (2019) 'genefilter: genefilter: methods for filtering genes from high-throughput experiments.' R package.

4. Alasoo, K. *et al.* (2018) 'Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response', *Nature Genetics*. Nature Publishing Group, 50(3), pp. 424–431. doi: 10.1038/s41588-018-0046-7.

5. Albert, F. W. and Kruglyak, L. (2015) 'The role of regulatory variation in complex traits and disease', *Nature Reviews Genetics*. doi: 10.1038/nrg3891.

6. Alexa A, R. J. (2019) 'topGO: Enrichment Analysis for Gene Ontology.' R package .

7. Alexeyev, M. *et al.* (2013) 'The maintenance of mitochondrial DNA integrity--critical analysis and update.', *Cold Spring Harbor perspectives in biology*. Cold Spring Harbor Laboratory Press, 5(5), p. a012641. doi: 10.1101/cshperspect.a012641.

8. Allaire, J. M. *et al.* (2018) 'The Intestinal Epithelium: Central Coordinator of Mucosal Immunity', *Trends in Immunology*. doi: 10.1016/j.it.2018.04.002.

9. Allis and Jenuwein (2016). 'The molecular hallmarks of epigenetic control., *Nature Reviews Genetics, 17(8), pp. 487-500.* doi: 10.1038/nrg.2016.59

10. Altshuler, D., Daly, M. J. and Lander, E. S. (2008) 'Genetic mapping in human disease', *Science*. doi: 10.1126/science.1156409.

11. Ananthakrishnan, A. N. (2015) 'Epidemiology and risk factors for IBD', *Nature Reviews Gastroenterology and Hepatology*. doi: 10.1038/nrgastro.2015.34.

12. Ananthakrishnan, A. N. *et al.* (2018) 'Environmental triggers in IBD: a review of progress and evidence', *Nature Reviews Gastroenterology & Hepatology*, 15(1), pp. 39–49. doi: 10.1038/nrgastro.2017.136.

13. Anders, S., Pyl, P. T. and Huber, W. (2015) 'HTSeq--a Python framework to work with

high-throughput sequencing data', *Bioinformatics*, 31(2), pp. 166–169. doi: 10.1093/bioinformatics/btu638.

14. Anderson, C. A. *et al.* (2011) 'Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47', *Nature Genetics*. doi: 10.1038/ng.764.

15. Andersson, R. *et al.* (2014) 'An atlas of active enhancers across human cell types and tissues', *Nature*, 507(7493), pp. 455–461. doi: 10.1038/nature12787.

16. Andrews, S. (2016b) *QC Fail Sequencing » RNA-Seq samples can be contaminated with DNA*. Available at: https://sequencing.qcfail.com/articles/rna-seq-samples-can-be-contaminated-with-dna/ (Accessed: 25 May 2019).

17. Andrews, S. (2019) *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 18 May 2019).

18. Arda, H. E. *et al.* (2018) 'A Chromatin Basis for Cell Lineage and Disease Risk in the Human Pancreas', *Cell Systems*, 7(3), pp. 310-322.e4. doi: 10.1016/j.cels.2018.07.007.

19. Astle, W. J. *et al.* (2016) 'The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease', *Cell*. doi: 10.1016/j.cell.2016.10.042.

20. Barko, P. C. *et al.* (2018) 'The Gastrointestinal Microbiome: A Review.', *Journal of veterinary internal medicine*. doi: 10.1111/jvim.14875.

21. Barnett, D. W. *et al.* (2011) 'BamTools: a C++ API and toolkit for analyzing and managing BAM files', *Bioinformatics*, 27(12), pp. 1691–1692. doi: 10.1093/bioinformatics/btr174.

22. Barrett, J. C. *et al.* (2009) 'Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes', *Nature Genetics*, 41(6), pp. 703–707. doi: 10.1038/ng.381.

23. Bauer, M. *et al*. (2017). 'Tobacco-smoking induced GPR15-expressing T cells in blood do not indicate pulmonary damage., *BMC Pulmonary Medicine*, 17(159). doi: 10.1186/s12890-017-0509-0

24. Baumgart, D. C. and Sandborn, W. J. (2007) 'Inflammatory bowel disease: clinical aspects and established and evolving therapies', *Lancet*. doi: 10.1016/S0140-6736(07)60751-X.

25. Benchimol, E. I. *et al.* (2015). 'Inflammatory bowel disease in immigrants to Canada and their children: a population-based cohort study., *American Journal of Gastroenterology*, 110(4), pp. 553–563. doi: 10.1038/ajg.2015.52

26. Berg, D. A. *et al.* (2019). 'A Common Embryonic Origin of Stem Cells Drives Developmental and Adult Neurogenesis., *Cell*, 177(3), pp. 654-668. doi: 10.1016/j.cell.2019.02.010

27. Bergenstrahle (2017) *Question: fdrtool - correction of p-values - course found on huber.embl.de*, *Bioconductor*. Available at: https://support.bioconductor.org/p/94020/ (Accessed: 20 May 2019).

28. Berlin, C. *et al.* (1993) 'Alpha 4 beta 7 integrin mediates lymphocyte binding to the mucosal vascular addressin MAdCAM-1.', *Cell*, 74(1), pp. 185–95. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7687523 (Accessed: 18 May 2019).

29. Biancheri, P. *et al.* (2014) 'Absence of a role for interleukin-13 in inflammatory bowel disease', *European Journal of Immunology*, 44(2), pp. 370–385. doi: 10.1002/eji.201343524.

30. Bioconductor Core Team (2019) 'TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s).' R package.

31. Bjerrum, J. T. *et al.* (2014) 'Transcriptional Analysis of Left-sided Colitis, Pancolitis, and Ulcerative Colitis-associated Dysplasia', *Inflammatory Bowel Diseases*, 20(12), pp. 2340–2352. doi: 10.1097/MIB.0000000000000235.

32. Bourgon, R., Gentleman, R. and Huber, W. (2010) 'Independent filtering increases detection power for high-throughput experiments', *Proceedings of the National*

Academy of Sciences, 107(21), pp. 9546–9551. doi: 10.1073/pnas.0914005107.

33. Bouzid, D. *et al.* (2013) 'Polymorphisms in the *IL2RA* and *IL2RB* Genes in Inflammatory Bowel Disease Risk', *Genetic Testing and Molecular Biomarkers*, 17(11), pp. 833–839. doi: 10.1089/gtmb.2013.0291.

34. Boyle, A. P. *et al.* (2008) 'High-resolution mapping and characterization of open chromatin across the genome.', *Cell*. NIH Public Access, 132(2), pp. 311–22. doi: 10.1016/j.cell.2007.12.014.

35. Brandtzaeg, P. *et al.* (1999) 'Regional specialization in the mucosal immune system: What happens in the microcompartments?', *Immunology Today*. doi: 10.1016/S0167-5699(98)01413-3.

36. Brodland, G. W. (2015). 'How computational models can help unlock biological systems., *Seminars in Cell and Developmental Biology*, 47-48, pp. 62-73. doi: 10.1016/j.semcdb.2015.07.001

37. Buenrostro, J. D. *et al.* (2013) 'Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position', *Nature Methods*. doi: 10.1038/nmeth.2688.

38. Buenrostro, J. D. *et al.* (2015) 'ATAC-seq: A method for assaying chromatin accessibility genome-wide', *Current Protocols in Molecular Biology*. doi: 10.1002/0471142727.mb2129s109.

39. Burisch, J. *et al.* (2013) 'The burden of inflammatory bowel disease in Europe', *Journal of Crohn's and Colitis*. doi: 10.1016/j.crohns.2013.01.010.

40. Bush, W. S. and Moore, J. H. (2012) 'Chapter 11: Genome-Wide Association Studies', *PLoS Computational Biology*. doi: 10.1371/journal.pcbi.1002822.

41. Bysani, M. *et al.* (2019). 'ATAC-seq reveals alterations in open chromatin in pancreatic islets from subjects with type 2 diabetes., *Scientific Reports,* 9(1), 7785. doi: 10.1038/s41598-019-44076-8

42. Cano-Gamez, E. and Trynka, G. (2020). 'From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases., *Frontiers in Genetics,* 11. doi: 10.3389/fgene.2020.00424.

43. Carlson (2019) 'org.Hs.eg.db: Genome wide annotation for Human.' R package.

44. Cha, H. *et al.* (2017) 'Increased susceptibility of IDH2-deficient mice to dextran sodium sulfate-induced colitis', *Redox Biology*, 13, pp. 32–38. doi: 10.1016/j.redox.2017.05.009.

45. Chakrabarti, A., Chen, A. W. and Varner, J. D. (2011) 'A review of the mammalian unfolded protein response', *Biotechnology and Bioengineering*, 108(12), pp. 2777–2793. doi: 10.1002/bit.23282.

46. Chang, C. (2019) *PLINK 1.9*. Available at: https://www.cog-genomics.org/plink/1.9/ (Accessed: 20 May 2019).

47. Chelakkot, C., Ghim, J. and Ryu, S. H. (2018) 'Mechanisms regulating intestinal barrier integrity and its pathological implications', *Experimental & molecular medicine*. doi: 10.1038/s12276-018-0126-x.

48. Chen, G. *et al.* (2014). 'Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data., *Human Molecular Genetics*, 23(17), pp. 4710–4720. doi: 10.1093/hmg/ddu174

49. Chun, S. *et al*. (2017). 'Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types., *Nature Genetics*, 49(4), pp. 600-605. doi: 10.1038/ng.3795.

50. Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*. doi: 10.1186/s13059-016-0881-8.

51. Corces, M. R. *et al.* (2016) 'Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution.', *Nature genetics*. NIH Public Access, 48(10), pp. 1193–203. doi: 10.1038/ng.3646.

52. Corces, M. R. *et al.* (2017) 'An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues.', *Nature methods*. NIH Public Access, 14(10),

pp. 959–962. doi: 10.1038/nmeth.4396.

53. Corces, M. R. *et al.* (2018). 'The chromatin accessibility landscape of primary human cancers., *Science*, 362(6413). doi: 10.1126/science.aav1898.

54. Costello, C. M. *et al.* (2005) 'Dissection of the Inflammatory Bowel Disease Transcriptome Using Genome-Wide cDNA Microarrays', *PLoS Medicine*. Edited by L. M. Sollid, 2(8), p. e199. doi: 10.1371/journal.pmed.0020199.

55. Crohn's and Colitis UK (no date) *IBD toolkit for GPs is now live | Crohn's & Colitis UK*. Available at: https://crohnsandcolitis.org.uk/news/ibd-toolkit-for-gps-launched (Accessed: 17 May 2019).

56. Danese, S. *et al.* (2015) 'Tralokinumab for moderate-to-severe UC: a randomised, double-blind, placebo-controlled, phase IIa study', *Gut*, 64(2), pp. 243–249. doi: 10.1136/gutjnl-2014-308004.

57. de Souza, H. S. P., Fiocchi, C. and Iliopoulos, D. (2017). 'The IBD interactome: an integrated view of aetiology, pathogenesis and therapy., *Nature Reviews Gastroenterology Hepatology*, 14(12), pp. 739-749. doi: 10.1038/nrgastro.2017.110

58. Degner, J. F. *et al.* (2012) 'DNase I sensitivity QTLs are a major determinant of human expression variation', *Nature*. Nature Publishing Group, 482(7385), pp. 390–394. doi: 10.1038/nature10808.

59. Dimas, A. S. *et al.* (2009) 'Common regulatory variation impacts gene expression in a cell type-dependent manner', *Science*. doi: 10.1126/science.1174148.

60. Ding, Y. *et al.* (2008) 'Beta-catenin stabilization extends regulatory T cell survival and induces anergy in nonregulatory T cells', *Nature Medicine*, 14(2), pp. 162–169. doi: 10.1038/nm1707.

61. Durinck, S. *et al.* (2005) 'BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis', *Bioinformatics*, 21(16), pp. 3439–3440. doi: 10.1093/bioinformatics/bti525.

62. Durinck, S. *et al.* (2009) 'Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt', *Nature Protocols*, 4(8), pp. 1184–1191.

doi: 10.1038/nprot.2009.97.

63. Ecker, J. R. *et al.* (2012) 'Genomics: ENCODE explained', *Nature*. doi: 10.1038/489052a.

64. Edwards, S. L. *et al.* (2013) 'Beyond GWASs: Illuminating the dark road from association to function', *American Journal of Human Genetics*. doi: 10.1016/j.ajhg.2013.10.012.

65. EMBL-EBI (2019a) *Expression Atlas*. Available at: https://www.ebi.ac.uk/gxa/home.

66. EMBL-EBI (2019b) *Single Cell Expression Atlas*. Available at: https://www.ebi.ac.uk/gxa/sc/home.

67. Ernst, J. *et al.* (2011) 'Mapping and analysis of chromatin state dynamics in nine human cell types', *Nature*. doi: 10.1038/nature09906.

68. Ewels, P. *et al.* (2016) 'MultiQC: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*. Narnia, 32(19), pp. 3047–3048. doi: 10.1093/bioinformatics/btw354.

69. Fairfax, B. P. *et al.* (2012) 'Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles', *Nature Genetics*. doi: 10.1038/ng.2205.

70. Fairfax, B. P. *et al.* (2014) 'Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression', *Science*. doi: 10.1126/science.1246949.

71. Farh, K. K. H. *et al.* (2015) 'Genetic and epigenetic fine mapping of causal autoimmune disease variants', *Nature*. doi: 10.1038/nature13835.

72. Fatahzadeh, M. (2009) 'Inflammatory bowel disease', *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*. Mosby, 108(5), pp. e1–e10. doi: 10.1016/J.TRIPLEO.2009.07.035.

73. Feagan, B. G. *et al.* (2013) 'Vedolizumab as Induction and Maintenance Therapy for Ulcerative Colitis', *New England Journal of Medicine*. Massachusetts Medical Society , 369(8), pp. 699–710. doi: 10.1056/NEJMoa1215734.

74. Felsenfeld, G. *et al.* (1996). 'Chromatin structure and gene expression., *PNAS*, 93(18), pp. 9384–9388. doi: 10.1073/pnas.93.18.9384.

75. Fields, P. E., Kim, S. T. and Flavell, R. A. (2002) 'Cutting Edge: Changes in Histone Acetylation at the IL-4 and IFN- Loci Accompany Th1/Th2 Differentiation', *The Journal of Immunology*, 169(2), pp. 647–650. doi: 10.4049/jimmunol.169.2.647.

76. Finucane, H. K. *et al.* (2015). 'Partitioning heritability by functional annotation using genomewide association summary statistics., *Nature Genetics*, 47(11), pp. 1228–1235. doi: 10.1038/ng.3404.

77. Fischer, A. *et al.* (2016). 'Differential effects of α4β7 and GPR15 on homing of effector and regulatory T cells from patients with UC to the inflamed gut in vivo., *Gut*, 65(10), pp.1642-1664. doi: 10.1136/gutjnl-2015-310022

78. Ford, A. C., Moayyedi, P. and Hanauer, S. B. (2013) 'Ulcerative colitis.', *BMJ (Clinical research ed.)*. British Medical Journal Publishing Group, 346, p. f432. doi: 10.1136/bmj.f432.

79. Fujino, S. *et al.* (2003) 'Increased expression of interleukin 17 in inflammatory bowel disease.', *Gut*. BMJ Publishing Group, 52(1), pp. 65–70. doi: 10.1136/gut.52.1.65.

80. Fullard, J. F. *et al.* (2018). 'An atlas of chromatin accessibility in the adult human brain., *Genome Research*, 28(8), pp.1243-1252. doi: 10.1101/gr.232488.117.

81. Fuss, I. J. *et al.* (1996) 'Disparate CD4+ lamina propria (LP) lymphokine secretion profiles in inflammatory bowel disease. Crohn's disease LP cells manifest increased secretion of IFN-gamma, whereas ulcerative colitis LP cells manifest increased secretion of IL-5.', *Journal of immunology (Baltimore, Md. : 1950)*, 157(3), pp. 1261–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8757634 (Accessed: 19 May 2019).

82. Fuss, I. J. *et al.* (2004) 'Nonclassical CD1d-restricted NK T cells that produce IL-13 characterize an atypical Th2 response in ulcerative colitis', *Journal of Clinical Investigation*, 113(10), pp. 1490–1497. doi: 10.1172/JCI19836.

83. García-Alcalde, F. *et al.* (2012) 'Qualimap: evaluating next-generation sequencing

alignment data', *Bioinformatics*, 28(20), pp. 2678–2679. doi: 10.1093/bioinformatics/bts503.

84. Gaspar, J. M. (2019) *ATAC-seq Guidelines*. Available at: https://informatics.fas.harvard.edu/atac-seq-guidelines.html (Accessed: 20 May 2019).

85. Gate, R. E. *et al.* (2018) 'Genetic determinants of co-accessible chromatin regions in activated T cells across humans', *Nature Genetics*. Nature Publishing Group, 50(8), pp. 1140–1150. doi: 10.1038/s41588-018-0156-2.

86. Gautier, E. L. *et al.* (2012) 'Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages.', *Nature immunology*. NIH Public Access, 13(11), pp. 1118–28. doi: 10.1038/ni.2419.

87. GBD 2017 Inflammatory Bowel Disease Collaborators (2020) 'The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017', *The Lancet Gastroenterol Hepatology*, 5(1), pp. 17-30. doi: 10.1016/S2468-1253(19)30333-4

88. Génin, E. (2019). 'Missing heritability of complex diseases: case solved?, *Human Genetics*, 139(1), pp. 103-113. doi: 10.1007/s00439-019-02034-4.

89. Gerrits, A. *et al.* (2009). 'Expression Quantitative Trait Loci Are Highly Sensitive to Cellular Differentiation State., *PLoS Genetics*, 5(10), e1000692. doi: 10.1371/journal.pgen.1000692

90. Gerstein, M. B. *et al.* (2012) 'Architecture of the human regulatory network derived from ENCODE data', *Nature*. doi: 10.1038/nature11245.

91. Giambartolomei, C. *et al.* (2018). 'A Bayesian framework for multiple trait colocalization from summary association statistics., *Bioinformatics*, 34(15), pp. 2538-2545. doi: 10.1093/bioinformatics/bty147.

92. Ginestet, C. (2011) 'ggplot2: Elegant Graphics for Data Analysis', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. doi: 10.1111/j.1467-985x.2010.00676_9.x.

93. Girard, J.-P., Moussion, C. and Förster, R. (2012) 'HEVs, lymphatics and homeostatic immune cell trafficking in lymph nodes', *Nature Reviews Immunology*, 12(11), pp. 762–773. doi: 10.1038/nri3298.

94. Gontarz, P. *et al.* (2020). 'Comparison of differential accessibility analysis strategies for ATAC-seq data., *Scientific Reports*, 10(1), 10150. doi: 10.1038/s41598-020-66998-4

95. Gordon, H. *et al.* (2015) 'Heritability in inflammatory bowel disease: From the first twin study to genome-wide association studies', *Inflammatory Bowel Diseases*. doi: 10.1097/MIB.0000000000000393.

96. Van der Goten, J. *et al.* (2014) 'Integrated miRNA and mRNA Expression Profiling in Inflamed Colon of Patients with Ulcerative Colitis', *PLoS ONE*. Edited by M. Chamaillard, 9(12), p. e116117. doi: 10.1371/journal.pone.0116117.

97. Gramlich, T. and Petras, R. E. (2007) 'Pathology of inflammatory bowel disease', *Seminars in Pediatric Surgery*. W.B. Saunders, 16(3), pp. 154–163. doi: 10.1053/J.SEMPEDSURG.2007.04.005.

98. Griffith, M. *et al.* (2015) 'Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud', *PLoS Computational Biology*. doi: 10.1371/journal.pcbi.1004393.

99. Grubert, F. *et al.* (2015) 'Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions.', *Cell*. NIH Public Access, 162(5), pp. 1051–65. doi: 10.1016/j.cell.2015.07.048.

100. GTEx Consortium *et al.* (2017) 'Genetic effects on gene expression across human tissues', *Nature*, 550(7675), pp. 204–213. doi: 10.1038/nature24277.

101. Gu, W. *et al.* (2016) 'Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications', *Genome Biology*. BioMed Central, 17(1), p. 41. doi: 10.1186/s13059-016-0904-5.

102. Guastadisegni, M. C. *et al.* (2010) 'CBFA2T2 and C20orf112: two novel fusion partners of RUNX1 in acute myeloid leukemia', *Leukemia*, 24(8), pp. 1516–1519. doi:

10.1038/leu.2010.106.

103. Gwiggner, M. *et al.* (2018) 'MicroRNA-31 and MicroRNA-155 are overexpressed in ulcerative colitis and regulate IL-13 signaling by targeting interleukin 13 receptor α-1', *Genes*. doi: 10.3390/genes9020085.

104. Haberman, Y. *et al.* (2019) 'Ulcerative colitis mucosal transcriptomes reveal mitochondriopathy and personalized mechanisms underlying disease severity and treatment response.', *Nature communications*. Nature Publishing Group, 10(1), p. 38. doi: 10.1038/s41467-018-07841-3.

105. Hahne, F. and Ivanek, R. (2016) 'Visualizing Genomic Data Using Gviz and Bioconductor', in. Humana Press, New York, NY, pp. 335–351. doi: 10.1007/978-1-4939-3578-9_16.

106. Hampe, J. *et al.* (2007) 'A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1', *Nature Genetics*. doi: 10.1038/ng1954.

107. Hansen et al (2019) 'Rgraphviz: Provides plotting capabilities for R graph objects.' R package.

108. Heinz, S. *et al.* (2015) 'The selection and function of cell type-specific enhancers', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 16(3), pp. 144–154. doi: 10.1038/nrm3949.

109. Herbert, Z. T. *et al.* (2018) 'Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction', *BMC Genomics*. doi: 10.1186/s12864-018-4585-1.

110. Holmén, N. *et al.* (2006) 'Functional CD4+CD25high regulatory T cells are enriched in the colonic mucosa of patients with active ulcerative colitis and increase with disease activity', *Inflammatory Bowel Diseases*, 12(6), pp. 447–456. doi: 10.1097/00054725-200606000-00003.

111. Hrdlickova, R., Toloue, M. and Tian, B. (2017) 'RNA-Seq methods for transcriptome analysis', *Wiley Interdisciplinary Reviews: RNA*. doi: 10.1002/wrna.1364.

112. Hu, X. *et al.* (2011). 'Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets., *AJHG*, 89(4), pp. 496-506. doi: 10.1016/j.ajhg.2011.09.002

113. Huang, H. *et al.* (2017) 'Fine-mapping inflammatory bowel disease loci to single-variant resolution', *Nature*, 547(7662), pp. 173–178. doi: 10.1038/nature22969.

114. Hummon, A. B. *et al.* (2012) 'Systems-wide RNAi analysis of CASP8AP2/FLASH shows transcriptional deregulation of the replication-dependent histone genes and extensive effects on the transcriptome of colorectal cancer cells', *Molecular Cancer*, 11(1), p. 1. doi: 10.1186/1476-4598-11-1.

115. Imai, Y. *et al.* (1999) 'The CED-4-homologous protein FLASH is involved in Fas-mediated activation of caspase-8 during apoptosis', *Nature*, 398(6730), pp. 777–785. doi: 10.1038/19709.

116. Inoue, S. *et al.* (1999) 'Characterization of cytokine expression in the rectal mucosa of ulcerative colitis: correlation with disease activity', *The American Journal of Gastroenterology*, 94(9), pp. 2441–2446. doi: 10.1111/j.1572-0241.1999.01372.x.

117. International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) (2015) *IBDGenetics - Home page*. Available at: https://www.ibdgenetics.org/ (Accessed: 17 May 2019).

118. Ito, D. *et al.* (2015) 'mTOR Complex Signaling through the SEMA4A–Plexin B2 Axis Is Required for Optimal Activation and Differentiation of CD8 $^{+}$ T Cells', *The Journal of Immunology*, 195(3), pp. 934–943. doi: 10.4049/jimmunol.1403038.

119. Ito, D. and Kumanogoh, A. (2016) 'The role of Sema4A in angiogenesis, immune responses, carcinogenesis, and retinal systems.', *Cell adhesion & migration*. Taylor & Francis, 10(6), pp. 692–699. doi: 10.1080/19336918.2016.1215785.

120. Ivanov, I. I. *et al.* (2006) 'The Orphan Nuclear Receptor RORγt Directs the Differentiation Program of Proinflammatory IL-17+ T Helper Cells', *Cell*, 126(6), pp. 1121–1133. doi: 10.1016/j.cell.2006.07.035.

121. Iwasaki, A. and Kelsall, B. L. (2001) 'Unique Functions of CD11b+, CD8 +, and Double-

Negative Peyer's Patch Dendritic Cells', *The Journal of Immunology*, 166(8), pp. 4884–4890. doi: 10.4049/jimmunol.166.8.4884.

122. Jerby-Arnon, L. *et al.* (2018) 'A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade', *Cell*, 175(4), pp. 984-997.e24. doi: 10.1016/j.cell.2018.09.006.

123. Johansson, M. E. V. *et al.* (2008) 'The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria', *Proceedings of the National Academy of Sciences*, 105(39), pp. 15064–15069. doi: 10.1073/pnas.0803124105.

124. John, S. *et al.* (2011) 'Chromatin accessibility pre-determines glucocorticoid receptor binding patterns', *Nature Genetics*. Nature Publishing Group, 43(3), pp. 264–268. doi: 10.1038/ng.759.

125. John, S. *et al.* (2013) 'Genome-scale mapping of DNase I hypersensitivity', *Current Protocols in Molecular Biology*. doi: 10.1002/0471142727.mb2127s103.

126. Jostins, L. *et al.* (2012) 'Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease', *Nature*. doi: 10.1038/nature11582.

127. Kabakchiev, B. and Silverberg, M. S. (2013) 'Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine', *Gastroenterology*. doi: 10.1053/j.gastro.2013.03.001.

128. Kadivar, K. *et al.* (2004) 'Intestinal Interleukin-13 in Pediatric Inflammatory Bowel Disease Patients', *Inflammatory Bowel Diseases*, 10(5), pp. 593–598. doi: 10.1097/00054725-200409000-00014.

129. Kanyavuz, A. *et al.* (2019). 'Breaking the law: unconventional strategies for antibody diversification., *Nature Reviews Immunology,* 19, pp. 355-368. doi.org/10.1038/s41577-019-0126-7

130. Kaplan, G. G. and Ng, S. C. (2017) 'Understanding and Preventing the Global Increase of Inflammatory Bowel Disease', *Gastroenterology*. doi: 10.1053/j.gastro.2016.10.020.

131. Karayiannakis, A. J. *et al.* (1998) 'Expression of catenins and E-cadherin during

epithelial restitution in inflammatory bowel disease', *The Journal of Pathology*, 185(4), pp. 413–418. doi: 10.1002/(SICI)1096-9896(199808)185:4<413::AID-PATH125>3.0.CO;2-K.

132. Kaser, A. *et al.* (2008) 'XBP1 Links ER Stress to Intestinal Inflammation and Confers Genetic Risk for Human Inflammatory Bowel Disease', *Cell*, 134(5), pp. 743–756. doi: 10.1016/j.cell.2008.07.021.

133. Kawamata, N. *et al.* (2012) 'Dominant-negative mechanism of leukemogenic PAX5 fusions', *Oncogene*, 31(8), pp. 966–977. doi: 10.1038/onc.2011.291.

134. Keshav, S. and Culver, E. (2011) *Gastroenterology : Clinical Cases Uncovered.* John Wiley & Sons.

135. Kim, D., Langmead, B. and Salzberg, S. L. (2015) 'HISAT: a fast spliced aligner with low memory requirements', *Nature Methods*, 12(4), pp. 357–360. doi: 10.1038/nmeth.3317.

136. Kim, J.-H. *et al.* (2017) 'CREB coactivators CRTC2 and CRTC3 modulate bone marrow hematopoiesis.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(44), pp. 11739–11744. doi: 10.1073/pnas.1712616114.

137. Kim, S.V. *et al.* (2013). 'GPR15-mediated homing controls immune homeostasis in the large intestine mucosa., *Science,* 340(6139), pp. 1456–1459. doi: 10.1126/science.1237013

138. King, C. G. *et al.* (2006) 'TRAF6 is a T cell–intrinsic negative regulator required for the maintenance of immune homeostasis', *Nature Medicine*, 12(9), pp. 1088–1092. doi: 10.1038/nm1449.

139. Kizil, C. *et al.* (2009) 'Simplet controls cell proliferation and gene transcription during zebrafish caudal fin regeneration', *Developmental Biology*, 325(2), pp. 329–340. doi: 10.1016/j.ydbio.2008.09.032.

140. Kizil, C. *et al.* (2014) 'Simplet/Fam53b is required for Wnt signal transduction by regulating -catenin nuclear localization', *Development*, 141(18), pp. 3529–3539. doi:

10.1242/dev.108415.

141. Klaus, B. (2014) *Differential expression analysis of RNA-Seq data using DESeq2*. Available at: https://www.huber.embl.de/users/klaus/Teaching/DESeq2-Analysis.pdf (Accessed: 26 May 2019).

142. Klaus, B. *et al.* (2016) *Analysis of RNA-Seq data: gene-level exploratory analysis and differential expression*. Available at: https://www.huber.embl.de/users/klaus/Teaching/DESeq2Predoc2014.html#inspection-and-correction-of-pvalues (Accessed: 5 August 2018).

143. Klemm, S. L., Shipony, Z. and Greenleaf, W. J. (2019). 'Chromatin accessibility and the regulatory epigenome., *Nature Reviews Genetics*, 20(4), pp. 207-220. doi: 10.1038/s41576-018-0089-8

144. De Klerk, E., Den Dunnen, J. T. and 'T Hoen, P. A. C. (2014) 'RNA sequencing: From tag-based profiling to resolving complete transcript structure', *Cellular and Molecular Life Sciences*. doi: 10.1007/s00018-014-1637-9.

145. Klonowski, K. D. *et al.* (2004) 'Dynamics of blood-borne CD8 memory T cell migration in vivo', *Immunity*. doi: 10.1016/S1074-7613(04)00103-7.

146. Kobayashi, T. *et al.* (2008) 'IL23 differentially regulates the Th1/Th17 balance in ulcerative colitis and Crohn's disease', *Gut*, 57(12), pp. 1682–1689. doi: 10.1136/gut.2007.135053.

147. Kobayashi, T. *et al.* (2020). 'Ulcerative colitis., *Nature Reviews Disease Primers*, 6(1). doi: 10.1038/s41572-020-0205-x.

148. Koch, U. *et al.* (2008) 'Simultaneous loss of - and -catenin does not perturb hematopoiesis or lymphopoiesis', *Blood*, 111(1), pp. 160–164. doi: 10.1182/blood-2007-07-099754.

149. Kruglyak, L. (2008) 'The road to genome-wide association studies', *Nature Reviews Genetics*. doi: 10.1038/nrg2316.

150. Kumar, B. V., Connors, T. J. and Farber, D. L. (2018) 'Human T Cell Development, Localization, and Function throughout Life', *Immunity*. doi:

10.1016/j.immuni.2018.01.007.

151. Kumar, B. V *et al.* (2017) 'Human Tissue-Resident Memory T Cells Are Defined by Core Transcriptional and Functional Signatures in Lymphoid and Mucosal Sites', *Cell Reports*, 20. doi: 10.1016/j.celrep.2017.08.078.

152. Kundaje, A. *et al.* (2015a) 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518(7539), pp. 317–330. doi: 10.1038/nature14248.

153. de la Torre-Ubieta, L. *et al.* (2018) 'The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis', *Cell*, 172(1–2), pp. 289-304.e18. doi: 10.1016/j.cell.2017.12.014.

154. Lamas, B. *et al.* (2016) 'CARD9 impacts colitis by altering gut microbiota metabolism of tryptophan into aryl hydrocarbon receptor ligands', *Nature Medicine*, 22(6), pp. 598–605. doi: 10.1038/nm.4102.

155. Landt, S. G. *et al.* (2012) 'ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.', *Genome research*. Cold Spring Harbor Laboratory Press, 22(9), pp. 1813–31. doi: 10.1101/gr.136184.111.

156. de Lange, K. M. *et al.* (2017) 'Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease', *Nature Genetics*, 49(2), pp. 256–261. doi: 10.1038/ng.3760.

157. Lawrence, M. *et al.* (2013) 'Software for Computing and Annotating Genomic Ranges', *PLoS Computational Biology*. Edited by A. Prlic. Public Library of Science, 9(8), p. e1003118. doi: 10.1371/journal.pcbi.1003118.

158. Lawrence, M., Daujat, S. and Schneider, R. (2016) 'Lateral Thinking: How Histone Modifications Regulate Gene Expression', *Trends in Genetics*, 32(1), pp. 42–56. doi: 10.1016/j.tig.2015.10.007.

159. Lee, C.-K. *et al.* (2004) 'Evidence for nucleosome depletion at active regulatory regions genome-wide', *Nature Genetics*, 36(8), pp. 900–905. doi: 10.1038/ng1400.

160. Lee, J. H. *et al.* (2007) 'Regulation of Ionizing Radiation-induced Apoptosis by Mitochondrial NADP $^{+}$ -dependent Isocitrate Dehydrogenase', *Journal of Biological*

Chemistry, 282(18), pp. 13385–13394. doi: 10.1074/jbc.M700303200.

161. Lee, M. N. *et al.* (2014) 'Common genetic variants modulate pathogen-sensing responses in human dendritic cells', *Science*. doi: 10.1126/science.1246980.

162. Lefrançois, L. *et al.* (1999) 'The Role of β7 Integrins in CD8 T Cell Trafficking During an Antiviral Immune Response', *The Journal of Experimental Medicine*, 189(10), pp. 1631–1638. doi: 10.1084/jem.189.10.1631.

163. Lehle, A. S. *et al.* (2019) 'Intestinal Inflammation and Dysregulated Immunity in Patients With Inherited Caspase-8 Deficiency', *Gastroenterology*, 156(1), pp. 275–278. doi: 10.1053/j.gastro.2018.09.041.

164. Lerner, A. and Matthias, T. (2015). 'Changes in intestinal tight junction permeability associated with industrial food additives explain the rising incidence of autoimmune disease., *Autoimmunity Rev.* 14(6), pp. 479–489. doi: 10.1016/j.autrev.2015.01.009

165. Levine, S. J. and Burakoff, R. (2011). 'Extraintestinal Manifestations of Inflammatory Bowel Disease., *Gastroenterology Hepatology*, 7(4), pp. 235-241.

166. Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

167. Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

168. Li, X. *et al.* (2011). 'Risk of inflammatory bowel disease in first- and second-generation immigrants in Sweden: a nationwide follow-up study., *Infammatory Bowel Disease*, 17(8), pp. 1784-1791. doi: 10.1002/ibd.21535

169. Liao, Y., Smyth, G. K. and Shi, W. (2014) 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics*, 30(7), pp. 923–930. doi: 10.1093/bioinformatics/btt656.

170. Lie, P. P. Y., Cheng, C. Y. and Mruk, D. D. (2011) 'The biology of the desmosome-like junction a versatile anchoring junction and signal transducer in the seminiferous epithelium.', *International review of cell and molecular biology*. NIH Public Access,

286, pp. 223–69. doi: 10.1016/B978-0-12-385859-7.00005-7.

171. Liu, J. Z. *et al.* (2015) 'Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations', *Nature Genetics*. doi: 10.1038/ng.3359.

172. Locke, A. E. *et al.* (2015) 'Genetic studies of body mass index yield new insights for obesity biology.', *Nature*. NIH Public Access, 518(7538), pp. 197–206. doi: 10.1038/nature14177.

173. Lodish, H. *et al.* (2000) *Molecular Cell Biology, 4th edition*. 4th edn. W.H. Freeman.

174. Loftus, E. V (2004) 'Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences.', *Gastroenterology*. doi: 10.1053/j.gastro.2004.01.063.

175. Lonsdale, J. *et al.* (2013) 'The Genotype-Tissue Expression (GTEx) project', *Nature Genetics*, 45(6), pp. 580–585. doi: 10.1038/ng.2653.

176. Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

177. Ludwig, L. S. *et al.* (2019). 'Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis', *Cell Reports*, 27(11), pp. 3228-3240.e7. doi: 10.1016/j.celrep.2019.05.046.

178. Luo, Y. *et al.* (2017) 'Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7.', *Nature genetics*. Europe PMC Funders, 49(2), pp. 186–192. doi: 10.1038/ng.3761.

179. Madrigal, P. (2015) 'On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions.', *Frontiers in bioengineering and biotechnology*. Frontiers Media SA, 3, p. 144. doi: 10.3389/fbioe.2015.00144.

180. Mangan, P. R. *et al.* (2006) 'Transforming growth factor-β induces development of the TH17 lineage', *Nature*, 441(7090), pp. 231–234. doi: 10.1038/nature04754.

181. Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), p. 10. doi: 10.14806/ej.17.1.200.

182. Mashima, R. *et al.* (2005) 'FLN29, a Novel Interferon- and LPS-inducible Gene Acting as a Negative Regulator of Toll-like Receptor Signaling', *Journal of Biological Chemistry*, 280(50), pp. 41289–41297. doi: 10.1074/jbc.M508221200.

183. Masopust, D. *et al.* (2010) 'Dynamic T cell migration program provides resident memory within intestinal epithelium', *The Journal of Experimental Medicine*. doi: 10.1084/jem.20090858.

184. Mathelier, A., Shi, W. and Wasserman, W. W. (2015) 'Identification of altered cis-regulatory elements in human disease', *Trends in Genetics*. doi: 10.1016/j.tig.2014.12.003.

185. Maurano, M. T. *et al.* (2012) 'Systematic localization of common disease-associated variation in regulatory DNA', *Science*. doi: 10.1126/science.1222794.

186. Mayassi, T. and Jabri, B. (2018) 'Human intraepithelial lymphocytes', *Mucosal Immunology*, 11(5), pp. 1281–1289. doi: 10.1038/s41385-018-0016-5.

187. McGovern, D. P. B. *et al.* (2010) 'Genome-wide association identifies multiple ulcerative colitis susceptibility loci', *Nature Genetics*. doi: 10.1038/ng.549.

188. Miles, C. and Wayne, M. (2008). 'Quantitative trait locus (QTL) analysis., *Nature Education*, 1(1), 208.

189. Miller, J. C. *et al.* (2012) 'Deciphering the transcriptional network of the dendritic cell lineage', *Nature Immunology*. doi: 10.1038/ni.2370.

190. Miller, R. K. *et al.* (2013) 'Beta-Catenin Versus the Other Armadillo Catenins', in *Progress in molecular biology and translational science*, pp. 387–407. doi: 10.1016/B978-0-12-394311-8.00017-0.

191. Moller, F. T. *et al.* (2015) 'Familial Risk of Inflammatory Bowel Disease: A Population-Based Cohort Study 1977–2011', *American Journal of Gastroenterology*, 110(4), pp. 564–571. doi: 10.1038/ajg.2015.50.

192. Molodecky, N. A. *et al.* (2012) 'Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review', *Gastroenterology*. doi: 10.1053/j.gastro.2011.10.001.

193. Momozawa, Y. *et al.* (2018) 'IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes', *Nature Communications*. Nature Publishing Group, 9(1), p. 2427. doi: 10.1038/s41467-018-04365-8.

194. Montefiori, L. *et al.* (2017) 'Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9', *Scientific Reports*. Nature Publishing Group, 7(1), p. 2451. doi: 10.1038/s41598-017-02547-w.

195. Moore, J. H., Asselbergs, F. W. and Williams, S. M. (2010). 'Bioinformatics challenges for genome-wide association studies., *Bioinformatics*, 26(4), pp. 445-455. doi: 10.1093/bioinformatics/btp713

196. Mozdiak, E., O'Malley, J. and Arasaradnam, R. (2015) 'Inflammatory bowel disease.', *BMJ*. British Medical Journal Publishing Group, 351, p. h4416. doi: 10.1136/bmj.h4416.

197. Mowat, A. M. *et al.* (2003) 'The role of dendritic cells in regulating mucosal immunity and tolerance.', *Novartis Foundation symposium*, 252, pp. 291–302; discussion 302-5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14609226 (Accessed: 2 June 2019).

198. Mowat, A. M. and Agace, W. W. (2014) 'Regional specialization within the intestinal immune system', *Nature Reviews Immunology*. doi: 10.1038/nri3738.

199. Mowat, A. M., Bain, C. C. and Mowat, A. M. (2014) 'Macrophages in intestinal homeostasis and inflammation.', *Immunological reviews*. doi: 10.1111/imr.12192.

200. Mullen, A. C. *et al.* (2001) 'Role of T-bet in Commitment of TH1 Cells Before IL-12-Dependent Selection', *Science*, 292(5523), pp. 1907–1910. doi: 10.1126/science.1059835.

201. Musunuru, K. *et al.* (2010). 'From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus., *Nature*, 466, pp. 714-719. doi: 10.1038/nature09266.

202. Nature Research Custom Media and Illumina, 2020. *Integrated multi-omics is more than sum of its parts.* Available at: https://www.nature.com/articles/d42473-020-00332-4?mvt=i&mvn=de6e28d463234e5eb76b3ce4ad2872ca&mvp=NA-NATUCOM-11239458&mvl=Fn-Homepage%20150%20%5BHome%20Layout%20-%20New%20Design%5D (Accessed: 2020)

203. Nemeth, Z. H. *et al.* (2017) 'Crohn's Disease and Ulcerative Colitis Show Unique Cytokine Profiles', *Cureus*, 9(4), p. e1177. doi: 10.7759/cureus.1177.

204. Nenci, A. *et al.* (2007) 'Epithelial NEMO links innate immunity to chronic intestinal inflammation.', *Nature*. doi: 10.1038/nature05698.

205. NICE (2019) *Ulcerative colitis: management*. Available at: https://www.nice.org.uk/guidance/ng130/resources/ulcerative-colitis-management-pdf-66141712632517 (Accessed: 01 Jun 2020).

206. Ng, S. C. *et al.* (2017) 'Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies', *The Lancet*, 390(10114), pp. 2769–2778. doi: 10.1016/S0140-6736(17)32448-0.

207. Nguyen, L. P. *et al.* (2015). 'Role and species-specific expression of colon T cell homing receptor GPR15 in colitis., *Nature Immunology*, 16(2), pp. 207-213. doi: 10.1038/ni.3079

208. NHGRI-EBI (2017) *GWAS Catalog*. Available at: https://www.ebi.ac.uk/gwas/ (Accessed: 17 May 2019).

209. Nicolae, D. L. *et al.* (2010) 'Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS', *PLoS Genetics*. Edited by G. Gibson, 6(4), p. e1000888. doi: 10.1371/journal.pgen.1000888.

210. NIESSNER, M. and VOLK, B. A. (2008) 'Altered Th1/Th2 cytokine profiles in the intestinal mucosa of patients with inflammatory bowel disease as assessed by quantitative reversed transcribed polymerase chain reaction (RT-PCR)', *Clinical & Experimental Immunology*, 101(3), pp. 428–435. doi: 10.1111/j.1365-2249.1995.tb03130.x.

211. NIH Roadmap Epigenomics Mapping Consortium (2010) *Roadmap Epigenomics Project - Home*. Available at: http://www.roadmapepigenomics.org/ (Accessed: 17 May 2019).

212. Okada, Y. *et al.* (2014) 'Genetics of rheumatoid arthritis contributes to biology and drug discovery', *Nature*, 506(7488), pp. 376–381. doi: 10.1038/nature12873.

213. Okou, D. T. *et al.* (2014) 'Exome Sequencing Identifies a Novel FOXP3 Mutation in a 2-Generation Family With Inflammatory Bowel Disease', *Journal of Pediatric Gastroenterology and Nutrition*, 58(5), pp. 561–568. doi: 10.1097/MPG.0000000000000302.

214. Okumura, R. and Takeda, K. (2017) 'Roles of intestinal epithelial cells in the maintenance of gut homeostasis', *Experimental & molecular medicine*. doi: 10.1038/emm.2017.20.

215. Olén, O. *et al*. (2020). 'Colorectal cancer in ulcerative colitis: a Scandinavian population-based cohort study., *The Lancet*, 395(10218), pp.123-131. doi: 10.1016/S0140-6736(19)32545-0.

216. Ott, J., Kamatani, Y. and Lathrop, M. (2011) 'Family-based designs for genome-wide association studies', *Nature Reviews Genetics*. doi: 10.1038/nrg2989.

217. Ou, J. *et al.* (2018) 'ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data', *BMC Genomics*. BioMed Central, 19(1), p. 169. doi: 10.1186/s12864-018-4559-3.

218. Ouyang, W., Kolls, J. K. and Zheng, Y. (2008) 'The Biological Functions of T Helper 17 Cell Effector Cytokines in Inflammation', *Immunity*, 28(4), pp. 454–467. doi: 10.1016/j.immuni.2008.03.004.

219. Pagès, Carlson and Falcon, L. (2019) 'AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor'. R package.

220. Parkes, M. *et al.* (2007) 'Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility', *Nature Genetics*, 39(7), pp. 830–832. doi: 10.1038/ng2061.

221. Parkes, M. *et al.* (2013) 'Genetic insights into common pathways and complex relationships among immune-mediated diseases', *Nature Reviews Genetics*. doi: 10.1038/nrg3502.

222. Patro, R., Mount, S. M. and Kingsford, C. (2014) 'Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms', *Nature Biotechnology*, 32(5), pp. 462–464. doi: 10.1038/nbt.2862.

223. Peloquin, J. M. *et al.* (2016) 'Characterization of candidate genes in inflammatory bowel disease–associated risk loci', *JCI Insight*. doi: 10.1172/jci.insight.87899.

224. Peterson, L. W. and Artis, D. (2014) 'Intestinal epithelial cells: Regulators of barrier function and immune homeostasis', *Nature Reviews Immunology*. doi: 10.1038/nri3608.

225. Pidasheva, S. *et al.* (2011) 'Functional Studies on the IBD Susceptibility Gene IL23R Implicate Reduced Receptor Function in the Protective Genetic Variant R381Q', *PLoS ONE*. Edited by S. K. Ahuja, 6(10), p. e25038. doi: 10.1371/journal.pone.0025038.

226. Pitule, P. *et al.* (2013) 'Differential expression and prognostic role of selected genes in colorectal cancer patients.', *Anticancer research*, 33(11), pp. 4855–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24222123 (Accessed: 3 June 2019).

227. Planell, N. *et al.* (2013) 'Transcriptional analysis of the intestinal mucosa of patients with ulcerative colitis in remission reveals lasting epithelial cell alterations', *Gut*, 62(7), pp. 967–976. doi: 10.1136/gutjnl-2012-303333.

228. Qu, K. *et al.* (2015) 'Individuality and variation of personal regulomes in primary human T cells.', *Cell systems*. NIH Public Access, 1(1), pp. 51–61. doi: 10.1016/j.cels.2015.06.003.

229. Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features.', *Bioinformatics (Oxford, England)*. Oxford University Press, 26(6), pp. 841–2. doi: 10.1093/bioinformatics/btq033.

230. Raine, T. *et al.* (2015) 'Generation of primary human intestinal T cell transcriptomes reveals differential expression at genetic risk loci for immune-mediated disease', *Gut*.

doi: 10.1136/gutjnl-2013-306657.

231. Reed, K. K. and Wickham, R. (2009) 'Review of the Gastrointestinal Tract: From Macro to Micro', *Seminars in Oncology Nursing*. doi: 10.1016/j.soncn.2008.10.002.

232. Reimold, A. M. *et al.* (2001) 'Plasma cell differentiation requires the transcription factor XBP-1', *Nature*, 412(6844), pp. 300–307. doi: 10.1038/35085509.

233. Reinisch, W. *et al.* (2015) 'Anrukinzumab, an anti-interleukin 13 monoclonal antibody, in active UC: efficacy and safety from a phase IIa randomised multicentre study', *Gut*, 64(6), pp. 894–900. doi: 10.1136/gutjnl-2014-308337.

234. Reznikoff, W. S. (2008) 'Transposon Tn *5*', *Annual Review of Genetics*. Annual Reviews , 42(1), pp. 269–286. doi: 10.1146/annurev.genet.42.110807.091656.

235. Riethoven, J.-J. M. (2010) 'Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators', in. doi: 10.1007/978-1-60761-854-6_3.

236. Rivas, M. A. *et al.* (2011a) 'Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease', *Nature Genetics*, 43(11), pp. 1066–1073. doi: 10.1038/ng.952.

237. Robinson, D. (2014) *How to interpret a p-value histogram – Variance Explained*. Available at: http://varianceexplained.org/statistics/interpreting-pvalue-histogram/ (Accessed: 19 May 2019).

238. Ross-Innes, C. S. *et al.* (2012) 'Differential oestrogen receptor binding is associated with clinical outcome in breast cancer', *Nature*. Nature Publishing Group, 481(7381), pp. 389–393. doi: 10.1038/nature10730.

239. Sanada, T. *et al.* (2008) 'FLN29 Deficiency Reveals Its Negative Regulatory Role in the Toll-like Receptor (TLR) and Retinoic Acid-inducible Gene I (RIG-I)-like Helicase Signaling Pathway', *Journal of Biological Chemistry*, 283(49), pp. 33858–33864. doi: 10.1074/jbc.M806923200.

240. Sandborn, W. J. *et al.* (2013) 'Vedolizumab as Induction and Maintenance Therapy for Crohn's Disease', *New England Journal of Medicine*. Massachusetts Medical Society , 369(8), pp. 711–721. doi: 10.1056/NEJMoa1215739.

241. Sandborn, W. J. *et al.* (2016) 'Ozanimod Induction and Maintenance Treatment for Ulcerative Colitis', *New England Journal of Medicine*. Massachusetts Medical Society, 374(18), pp. 1754–1762. doi: 10.1056/NEJMoa1513248.

242. Santos, M. P. C., Gomes, C. and Torres, J. (2018) 'Familial and ethnic risk in inflammatory bowel disease.', *Annals of gastroenterology*. The Hellenic Society of Gastroenterology, 31(1), pp. 14–23. doi: 10.20524/aog.2017.0208.

243. Sathaliyawala, T. *et al.* (2013) 'Distribution and Compartmentalization of Human Circulating and Tissue-Resident Memory T Cell Subsets', *Immunity*. doi: 10.1016/j.immuni.2012.09.020.

244. Schägger, H. *et al.* (1995) 'Ubiquinol-cytochrome-c reductase from human and bovine mitochondria.', *Methods in enzymology*, 260, pp. 82–96. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8592474 (Accessed: 3 June 2019).

245. Scharer, C. D. *et al.* (2016). 'ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells., *Scientific Reports*, 6, 27030. doi: 10.1038/srep27030

246. Schaub, M. A. *et al.* (2012) 'Linking disease associations with regulatory information in the human genome', *Genome Research*. doi: 10.1101/gr.136127.111.

247. Schmidt, E. M. *et al.* (2015). 'GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach., *Bioinformatics*, 31 (16), pp. 2601–2606. doi: 10.1093/bioinformatics/btv201.

248. Schuierer, S. *et al.* (2017) 'A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples', *BMC Genomics*. doi: 10.1186/s12864-017-3827-y.

249. Scott-Browne, J. P. *et al.* (2016) 'Dynamic Changes in Chromatin Accessibility Occur in CD8 + T Cells Responding to Viral Infection', *Immunity*, 45(6), pp. 1327–1340. doi: 10.1016/j.immuni.2016.10.028.

250. Sheridan, B. S. and Lefrançois, L. (2011) 'Regional and mucosal memory T cells', *Nature Immunology*. doi: 10.1038/ni.2029.

251. Silverberg, M. S. *et al.* (2009) 'Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study', *Nature Genetics*. doi: 10.1038/ng.275.

252. Singh, T. *et al.* (2015) 'Characterization of expression quantitative trait loci in the human colon', *Inflammatory Bowel Diseases*. doi: 10.1097/MIB.0000000000000265.

253. Sleiman, P. M. A. *et al.* (2010) 'Variants of *DENND1B* Associated with Asthma in Children', *New England Journal of Medicine*, 362(1), pp. 36–44. doi: 10.1056/NEJMoa0901867.

254. Smillie, C. S. *et al.* (2018) 'Rewiring of the cellular and inter-cellular landscape of the human colon during ulcerative colitis', *bioRxiv*. Cold Spring Harbor Laboratory, p. 455451. doi: 10.1101/455451.

255. Sohn, J. J. *et al.* (2012) 'Macrophages, Nitric Oxide and microRNAs Are Associated with DNA Damage Response Pathway and Senescence in Inflammatory Bowel Disease', *PLoS ONE*. Edited by B. Foligne. Public Library of Science, 7(9), p. e44156. doi: 10.1371/journal.pone.0044156.

256. Soneson, C., Love, M. I. and Robinson, M. D. (2016) 'Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences', *F1000Research*, 4, p. 1521. doi: 10.12688/f1000research.7563.2.

257. Soskic, B. *et al.* (2019) 'Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases', *bioRxiv*. Cold Spring Harbor Laboratory, p. 566810. doi: 10.1101/566810.

258. Spain, S. L. and Barrett, J. C. (2015) 'Strategies for fine-mapping complex traits', *Human Molecular Genetics*. doi: 10.1093/hmg/ddv260.

259. Spencer, J. and Sollid, L. M. (2016) 'The human intestinal B-cell response', *Mucosal Immunology*, 9(5), pp. 1113–1124. doi: 10.1038/mi.2016.59.

260. Spindel, O. N., World, C. and Berk, B. C. (2012) 'Thioredoxin Interacting Protein: Redox Dependent and Independent Regulatory Mechanisms', *Antioxidants & Redox Signaling*, 16(6), pp. 587–596. doi: 10.1089/ars.2011.4137.

261. Stark, R. and Brown, G. D. (2011) 'DiffBind: differential binding analysis of ChIP-Seq peak data'.

262. Strimmer, K. (2008) 'fdrtool: a versatile R package for estimating local and tail area-based false discovery rates', *Bioinformatics*, 24(12), pp. 1461–1462. doi: 10.1093/bioinformatics/btn209.

263. Strober, W. and Watanabe, T. (2011) 'NOD2, an intracellular innate immune sensor involved in host defense and Crohn's disease', *Mucosal Immunology*. doi: 10.1038/mi.2011.29.

264. Sun, M. *et al.* (2017) 'Cbx3/HP1γ deficiency confers enhanced tumor-killing capacity on CD8+ T cells', *Scientific Reports*. Nature Publishing Group, 7(1), p. 42888. doi: 10.1038/srep42888.

265. Svaren, J. *et al.* (1994) 'Analysis of the competition between nucleosome formation and transcription factor binding.', *The Journal of biological chemistry*, 269(12), pp. 9335–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8132673 (Accessed: 3 June 2019).

266. Szabo, S. J. *et al.* (2000) 'A novel transcription factor, T-bet, directs Th1 lineage commitment.', *Cell*, 100(6), pp. 655–69. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10761931 (Accessed: 2 June 2019).

267. Sýkora, S. *et al.* (2018) 'Current global trends in the incidence of pediatric-onset inflammatory bowel disease', *World Journal Gastroenterol*, 24(25), pp. 2741-2763. doi: 10.3748/wjg.v24.i25.2741

268. Takahashi, Y. *et al.* (2007) 'Decreased expression of thioredoxin interacting protein mRNA in inflamed colonic mucosa in patients with ulcerative colitis.', *Oncology reports*, 18(3), pp. 531–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17671698 (Accessed: 20 May 2019).

269. Takara Bio Inc. (2018) *Pushing the limit: a complete solution for generating stranded RNA-seq libraries from picogram inputs of total mammalian RNA*. Available at: https://www.takarabio.com/learning-centers/next-generation-sequencing/technical-notes/rna-seq/stranded-libraries-from-picogram-input-total-

rna-(v1) (Accessed: 25 May 2019).

270. Takara Bio USA (no date) *SMARTer® Stranded Total RNA-Seq Kit-Pico Input Mammalian User Manual SMARTer Stranded Total RNA-Seq Kit-Pico Input Mammalian User Manual*. Available at: https://www.takarabio.com/assets/documents/User Manual/SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian User Manual_112216.pdf (Accessed: 17 May 2019).

271. Takata, A. *et al.* (2010) 'Behavioral and gene expression analyses in heterozygous XBP1 knockout mice: Possible contribution of chromosome 11qA1 locus to prepulse inhibition', *Neuroscience Research*, 68(3), pp. 250–255. doi: 10.1016/j.neures.2010.07.2042.

272. Taman, H. *et al.* (2018) 'Transcriptomic Landscape of Treatment—Naïve Ulcerative Colitis', *Journal of Crohn's and Colitis*, 12(3), pp. 327–336. doi: 10.1093/ecco-jcc/jjx139.

273. Targan, S. R. *et al.* (2007) 'Natalizumab for the Treatment of Active Crohn's Disease: Results of the ENCORE Trial', *Gastroenterology*, 132(5), pp. 1672–1683. doi: 10.1053/j.gastro.2007.03.024.

274. Than, T. A. *et al.* (2011) 'Role of cAMP-responsive Element-binding Protein (CREB)-regulated Transcription Coactivator 3 (CRTC3) in the Initiation of Mitochondrial Biogenesis and Stress Response in Liver Cells', *Journal of Biological Chemistry*, 286(25), pp. 22047–22054. doi: 10.1074/jbc.M111.240481.

275. The Broad Institute (2019) *Singel Cell BETA Portal*. Available at: https://portals.broadinstitute.org/single_cell.

276. The ENCODE Project Consortium (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*. Nature Publishing Group, 489(7414), pp. 57–74. doi: 10.1038/nature11247.

277. The IBD Standards Group (2013) *IBD Standards Standards for the Healthcare of People who have Inflammatory Bowel Disease (IBD) 2013 Update*. Available at: http://s3-eu-west-1.amazonaws.com/files.crohnsandcolitis.org.uk/Publications/PPR/ibd-

standards.pdf (Accessed: 17 May 2019).

278. Thiel, G. *et al.* (2000) 'The human transcriptional repressor protein NAB1: expression and biological activity.', *Biochimica et biophysica acta*, 1493(3), pp. 289–301. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11018254 (Accessed: 3 June 2019).

279. Thome, J. J. C. *et al.* (2016) 'Early-life compartmentalization of human T cell differentiation and regulatory function in mucosal and lymphoid tissues', *Nature Medicine*, 22(1), pp. 72–77. doi: 10.1038/nm.4008.

280. Thurman, R. E. *et al.* (2012) 'The accessible chromatin landscape of the human genome', *Nature*. doi: 10.1038/nature11232.

281. Trapnell, C. *et al.* (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature Biotechnology*, 28(5), pp. 511–515. doi: 10.1038/nbt.1621.

282. Turpin, W. *et al.* (2018). 'Determinants of IBD Heritability: Genes, Bugs, and More., *Infammatory Bowel Disease*, 24(6), pp. 1133-1148. doi: 10.1093/ibd/izy085

283. Vadasz, Z. *et al.* (2015) 'The Involvement of Immune Semaphorins in the Pathogenesis of Inflammatory Bowel Diseases (IBDs)', *PLOS ONE*. Edited by D. L. Boone. Public Library of Science, 10(5), p. e0125860. doi: 10.1371/journal.pone.0125860.

284. Vainer, B. *et al.* (2000) 'COLONIC EXPRESSION AND SYNTHESIS OF INTERLEUKIN 13 AND INTERLEUKIN 15 IN INFLAMMATORY BOWEL DISEASE', *Cytokine*, 12(10), pp. 1531–1536. doi: 10.1006/cyto.2000.0744.

285. Veldhoen, M. *et al.* (2008) 'Transforming growth factor-β "reprograms" the differentiation of T helper 2 cells and promotes an interleukin 9–producing subset', *Nature Immunology*, 9(12), pp. 1341–1346. doi: 10.1038/ni.1659.

286. Vermeire, S. *et al.* (2014) 'Etrolizumab as induction therapy for ulcerative colitis: a randomised, controlled, phase 2 trial.', *Lancet (London, England)*. Elsevier, 384(9940), pp. 309–18. doi: 10.1016/S0140-6736(14)60661-9.

287. Vermeire, S. *et al.* (2017) 'Anti-MAdCAM antibody (PF-00547659) for ulcerative colitis

(TURANDOT): a phase 2, randomised, double-blind, placebo-controlled trial', *The Lancet*, 390(10090), pp. 135–144. doi: 10.1016/S0140-6736(17)30930-3.

288. Villarino, A. V. *et al.* (2015) 'Mechanisms of Jak/STAT Signaling in Immunity and Disease', *The Journal of Immunology*, 194(1), pp. 21–27. doi: 10.4049/jimmunol.1401867.

289. Wang, W. Y. S. *et al.* (2005) 'Genome-wide association studies: Theoretical and practical concerns', *Nature Reviews Genetics*. doi: 10.1038/nrg1522.

290. Weedon, M. N. *et al.* (2014) 'Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis', *Nature Genetics*, 46(1), pp. 61–64. doi: 10.1038/ng.2826.

291. Wellcome Trust Case Control Consortium (2007) 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, 447(7145), pp. 661–678. doi: 10.1038/nature05911.

292. Wen, X., Pique-Regi, R. and Luca, F. (2017). 'Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization., *PLoS Genetics*, 13(3), e1006646. doi: 10.1371/journal.pgen.1006646.

293. Wirtz, S. *et al.* (2017) 'Chemically induced mouse models of acute and chronic intestinal inflammation', *Nature Protocols*. doi: 10.1038/nprot.2017.044.

294. Wood, A. R. *et al.* (2014) 'Defining the role of common variation in the genomic and biological architecture of adult human height', *Nature Genetics*, 46(11), pp. 1173–1186. doi: 10.1038/ng.3097.

295. Wood, D. E. and Salzberg, S. L. (2014) 'Kraken: ultrafast metagenomic sequence classification using exact alignments', *Genome Biology*. BioMed Central, 15(3), p. R46. doi: 10.1186/gb-2014-15-3-r46.

296. Workman, J. L. and Kingston, R. E. (1992) 'Nucleosome core displacement in vitro via a metastable transcription factor-nucleosome complex.', *Science (New York, N.Y.)*, 258(5089), pp. 1780–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1465613 (Accessed: 3 June 2019).

297. Wu, J. *et al.* (2016) 'The landscape of accessible chromatin in mammalian preimplantation embryos', *Nature*, 534(7609), pp. 652–657. doi: 10.1038/nature18606.

298. Xue, F. *et al.* (2013) 'MiR-31 regulates interleukin 2 and kinase suppressor of ras 2 during T cell activation', *Genes and Immunity*. doi: 10.1038/gene.2012.58.

299. Yamada, T. *et al.* (2008) *Principles of clinical gastroenterology*. Wiley-Blackwell.

300. Yang, C.-W. *et al.* (2016) 'Regulation of T Cell Receptor Signaling by DENND1B in T H 2 Cells and Allergic Disease', *Cell*, 164(1–2), pp. 141–155. doi: 10.1016/j.cell.2015.11.052.

301. Yang, H. *et al.* (2012) 'IDH1 and IDH2 Mutations in Tumorigenesis: Mechanistic Insights and Clinical Perspectives', *Clinical Cancer Research*, 18(20), pp. 5562–5571. doi: 10.1158/1078-0432.CCR-12-1773.

302. Yoshimura, S. *et al.* (2010) 'Family-wide characterization of the DENN domain Rab GDP-GTP exchange factors', *The Journal of Cell Biology*, 191(2), pp. 367–381. doi: 10.1083/jcb.201008051.

303. Yu, G. *et al.* (2012) 'clusterProfiler: an R package for comparing biological themes among gene clusters.', *Omics : a journal of integrative biology*. Mary Ann Liebert, Inc., 16(5), pp. 284–7. doi: 10.1089/omi.2011.0118.

304. Yu, G. *et al.* (2015) 'DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis', *Bioinformatics*, 31(4), pp. 608–609. doi: 10.1093/bioinformatics/btu684.

305. Yu, G. and He, Q.-Y. (2016) 'ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization', *Molecular BioSystems*. The Royal Society of Chemistry, 12(2), pp. 477–479. doi: 10.1039/C5MB00663E.

306. Yu, G., Wang, L.-G. and He, Q.-Y. (2015) 'ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization', *Bioinformatics*, 31(14), pp. 2382–2383. doi: 10.1093/bioinformatics/btv145.

307. Yu, Q. T. *et al.* (2007) 'Expression and functional characterization of FOXP3+CD4+

regulatory T cells in ulcerative colitis', *Inflammatory Bowel Diseases*, 13(2), pp. 191–199. doi: 10.1002/ibd.20053.

308. Yu, W., Dittenhafer-Reed, K. E. and Denu, J. M. (2012) 'SIRT3 Protein Deacetylates Isocitrate Dehydrogenase 2 (IDH2) and Regulates Mitochondrial Redox Status', *Journal of Biological Chemistry*, 287(17), pp. 14078–14086. doi: 10.1074/jbc.M112.355206.

309. Zammarchi, I. *et al.* (2020) 'Elderly-onset vs adult-onset ulcerative colitis: a different natural history?', *BMC Gastroenterol,* 20(147). https://doi.org/10.1186/s12876-020-01296-x

310. Zhang, Y. *et al.* (2008) 'Model-based Analysis of ChIP-Seq (MACS)', *Genome Biology*, 9(9), p. R137. doi: 10.1186/gb-2008-9-9-r137.

311. Zhao, S. *et al*. (2018). 'RnaSeqSampleSize: real data based sample size estimation for RNA sequencing., *BMC Bioinformatics*, 9(1). doi: 10.1186/s12859-018-2191-5

312. Zhao, S. *et al.* (2018) 'Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: PolyA+ selection versus rRNA depletion', *Scientific Reports*. doi: 10.1038/s41598-018-23226-4.

313. Zhao, S., Guo, Y. and Shyr, Y. (2019) *KEGGprofile: An annotation and visualization package for multi-types and multi-groups expression data in KEGG pathway.* Available at: https://bioconductor.org/packages/release/bioc/html/KEGGprofile.html (Accessed: 17 May 2019).

314. Zhao, W. *et al.* (2014) 'Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling', *BMC Genomics*. doi: 10.1186/1471-2164-15-419.

315. Zhu, L. *et al.* (2017) 'IL-10 and IL-10 Receptor Mutations in Very Early Onset Inflammatory Bowel Disease.', *Gastroenterology research*. Elmer Press, 10(2), pp. 65–69. doi: 10.14740/gr740w.

Appendix 1 –RNA Sequencing Optimization

There is no gold standard for sequencing design. An optimal sequencing depth is highly dependent on the aims of each individual experiment (Conesa *et al.*, 2016). Sequencing depth in this case refers to the total number of sequenced reads in a sample.

Before venturing into a large-scale sequencing project, we first set out to determine the optimal sequencing depth, which could provide us with sufficient level of information to carry out the intended analysis. In addition, as it was our first-time constructing RNA seq libraries, we wanted to assess sample quality before further committing to high in cost sequencing experiment.

Initial sequencing parameters were selected based on recommendations from experts in the Medimmune Bioinformatics facility and Dr. Bergamaschi (Postdoctoral Fellow in Prof Ken Smith lab). Briefly, 13 RNA Seq libraries were pooled together, quality checked and run on NovaSeq 500 (75bp, PE, 30M reads/Sample) located in the Department of Biochemistry, University of Cambridge. Initial quality assessment and read mapping were performed by the MedImmune Bioinformatics facility using the bcbio open resource python toolkit.

Raw Sequencing Read Alignment

Sample alignment ranged from 1.7% to 81% of total reads sequenced, with median of 41% (Figure A1.1). Poor mapability together with GC content bias raised the question of possible contamination with other (non-human) genomes.

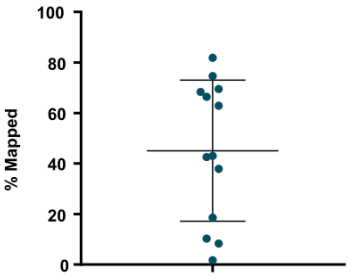


Figure A1.1 RAW RNA SEQUENCING READ ALIGNMENT TO THE HUMAN GENOME EXPRESSED AS PERCENTAGE OF TOTAL READS ($n_{sample} = 13$). *Each dot represents a sample. Error bar represents the mean and standard deviation.*

To test for potential impurities Kraken (a system for assigning taxonomic labels to short DNA fragments) was used. In short, after the low complexity region removal, unmapped reads were fragmented into the smaller fragments and passed to the Kraken to test for the contamination with mammalian, viral and bacterial genomes.

Only a small proportion (median = 19.8 %) of unmapped reds mapped to the contaminants selected, leaving the rest of unmapped reads unclassified (Figure A1.2).

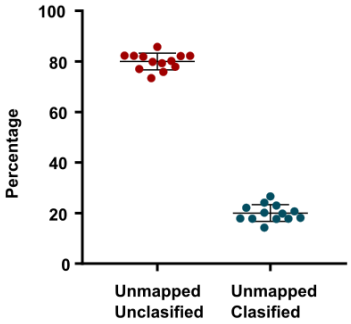


Figure A1.2 PERCENTAGE OF UNMAPPED READS THAT EITHER DID OR DID NOT REALIGN AGAINST ANY OF OTHER GENOMES TESTED ($n_{sample} = 13$). *Unmapped-Unclassified shows reads that failed any alignment, whereas Unmapped-Classified found their alignment after Kraken screening. Each dot represents a sample. Error bar represents the mean and standard deviation.*

The Unmapped-Classified reads mapped to human and bacterial genomes, but not viral or other mammalian genomes. Interestingly, the percentage of remapping strongly correlated with sample RIN (Figure A1.3), where increases in RNA quality showed strong negative correlation ($r = -0.775$, $p = 0.0028$) with mapping to bacterial

genome and strong, positive correlation (*r = 0.775, p = 0.0028*) with mapping to human genome.

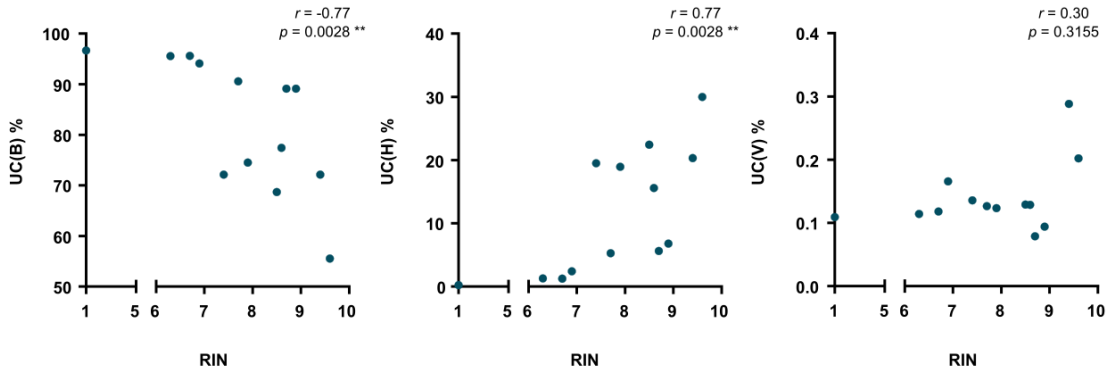


Figure A1.3 RELATIONSHIP BETWEEN THE RNA QUALITY AND PERCENTAGE OF READ RE-MAPPING TO THE VIRAL, BACTERIAL AND HUMAN GENOME (n $_{sample}$ = 13). *Relationship was quantified by Spearman's correlation. Each dot represents a sample. r - Spearman's rho; p - p-value; RIN - RNA integrity number; UC - Unmapped-Classified; B – Bacterial Genome; H – Human Genome; V – Viral Genome.*

Taken together, these observations suggested that most of the unmapped reads could not be attributable to the contamination with either viral or mammalian genetic material. Rather it seems that the samples with low RNA quality are more susceptible to the presence of bacterial genome.

Read Genomic Alignment

Analysis revealed that most of the mapped reads aligned to the intergenic regions (median = 63.8%) with the second largest group mapping to intronic regions (median = 25.35%) and only a small proportion of total aligned reads mapped to the exonic regions (median = 10.55%) (Figure A1.4).

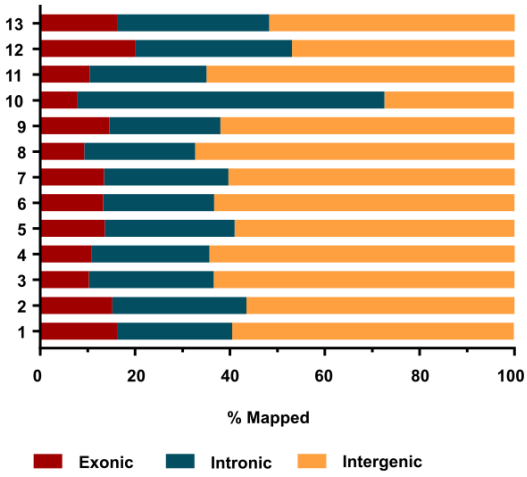


Figure A1.4 GENOMIC ALIGNMENT OF MAPPED READS. *Samples 1 to 13 are shown along y axis and the % of total mapped reads that fell on defined genomic region are shown amongst the x axis.*

We had two hypotheses to explain why a large proportion of reads mapped to the intergenic regions and not to the exons. We speculated that libraries could either have been contaminated with the gDNA or the rRNA depletion methodology used might be responsible for this unusual read distribution.

To exclude gDNA as a possible contaminant, RNA purity was re-assessed. First, we re-valuated all RNA sample associated electropherograms. None of gDNA characteristic peaks were seen in any of RNA samples. Second, we used Qubit benchtop fluorometer to accurately measure gDNA quantity in 7 of the 13 RNA samples used for library generation. All samples were below detection range (<10pg/ul).

At time of our Optimization experiments, Illumina released Ribo-Zero rRNA removal kits, which worked in similar manner to the SMARTer Stranded Total RNA seq Kit- Pico Input Mammalian kit we used for our RNA seq library construction. The supplementary information published along the Ribo-Zero rRNA removal kits showed that libraries constructed by rRNA depletion method shows significantly different genomic alignment that RNA Seq libraries made by conventional poly-A enrichment methods. Therefor we concluded that there was no error during library construction and the unusual alignment is due the new library generation technology.

The main finding from the first optimization experiment was that only a small fraction of reads aligns to the exonic regions and, hereby, sequencing depth of 30M is far too low for accurately determining changes in gene expression. Therefore, we decided to run a second optimization experiment increasing the sequencing depth to 200M PE reads/sample that allowed us to down sample and find the optimal sequencing depth (balance between the information needed and the associated costs). Finally, we selected sequencing depth of 110M PE reads/sample for our main experiment.

Appendix 2 – *In Silico* Predictions Of Individual RNA Seq Library Sequencing Performance

From optimization experiments we learned that some of the RNA seq libraries showed very poor sequencing performance. Early identification and exclusion of these libraries would be very beneficial from the financial perspective. Therefore, to predict which libraries will result in poor performance cell count, RIN, extracted RNA concentration and library insert size were correlated with the percentage of total reads mapped (Figure A2.1).

Both - RIN and extracted RNA concentration showed strong correlation (*r = 0.67, p =0.015* and *r = 0.819, p = 0.001*) with read alignment. Though cell number showed moderate correlation with mapping ability, the correlation coefficient was not significant (*r = 0.443, p = 0.13*). However, cell number exhibited robust correlated with both RIN (*r = 0.633, p = 0.023*) and extracted RNA concentration (*r = 0.575 , p = 0.043*) (Data not shown).

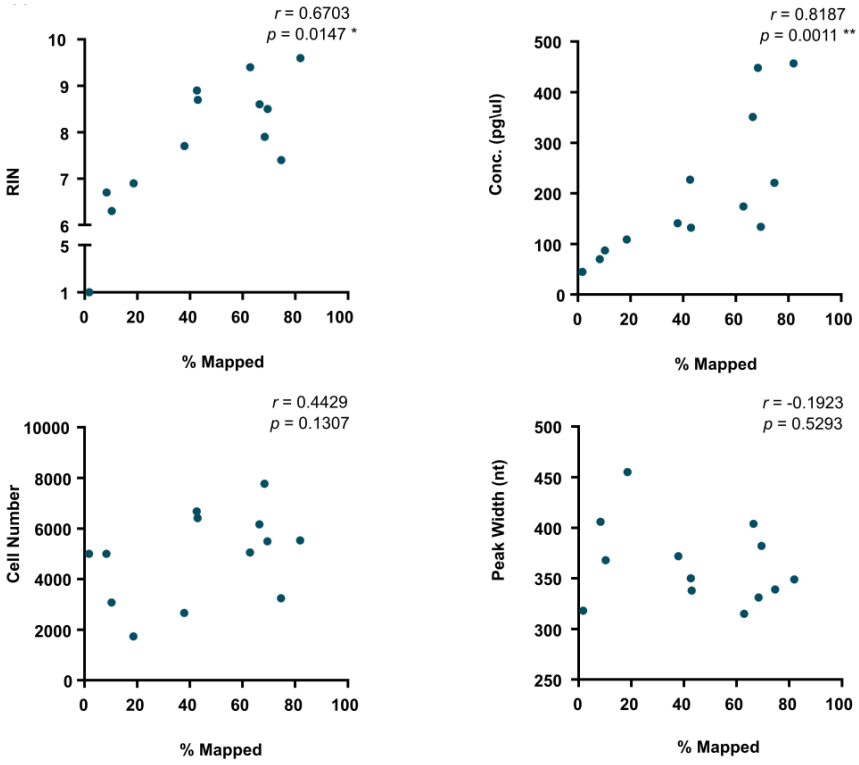


FIGURE CONTINUED IN NEXT PAGE

Figure A2.1 RELATIONSHIP BETWEEN THE RNA SEQUENCING LIBRARY ALIGNMENT AND EARLY SAMPLE/LIBRARY METRICS (n $_{sample}$ = 13). *Relationship was quantified by Spearman's correlation. Each dot represents a sample. r - Spearman's rho; p - p-value; RIN - RNA integrity number; Conc. – extracted RNA concentration.*

In summary, we identified that libraries generated from samples low in RNA quantity or quality resulted in poor alignment to the human genome and showed increased contamination with bacterial genome. It should be mentioned that all libraries were generated from the same amount of RNA, thus, the impact of initial RNA concentration on library quality was not anticipated. However, it might explain the bacterial contamination seen, as, most likely, low amounts of RNA in tubes was not enough to dilute out environmental contamination, which further was successfully amplified during the PCR amplification step. Therefore, it was decided to exclude samples with RIN <7 or concentration < 100pg/ul from further study. Both thresholds were selected after analysis of correlation scatter plots, keeping in mind that too stringent thresholding would reduce sample numbers and hence power.

Appendix 3 – Extended RNA Seq Data QC

In RNA and ATAC sequencing experiments large emphasis was put on the data QC. We believe that QC is a very important part of any experiment. It not only examines the presence of any technical artifacts, but also if general biological assumptions are satisfied.

Sequencing Library Quality Control

For initial quality assessment and read mapping the same pipeline as for both optimization runs was performed by the MedImmune Bioinformatics facility.

A median sequencing depth per library of 82.5M PE reads and 98.2M PE reads for sequencing runs 1 and 2 was achieved (Figure A3.1 A). In comparison to the optimization run, median read count mapped increased to 70.25M PE reads per library, with the lowest sample achieving 36.10M PE reads (Figure A3.1 B).

Figure A3.1 RNA LIBRARY SEQUENCING DEPTH AND ALIGNMENT *($n_{sample} = 92$). A. The number of PE reads each sample was sequenced to. B. Percentage of total sequenced reads per sample that mapped to the human genome. Each dot represents a sample. Error bar represents the mean and standard deviation. PE - Paired end sequencing; M – Million.*

Moreover, percentage of mapped reads moderately correlated with RIN number (*$r = 0.428$, $p = 0.0001$*), and cell count (*$r = 0.4311$, $p = 0.0001$*), but kept very strong correlation with RNA concentration (pg/ul) (*$r = 0.7396$, $p = 0.0001$*) (Figure A3.2).

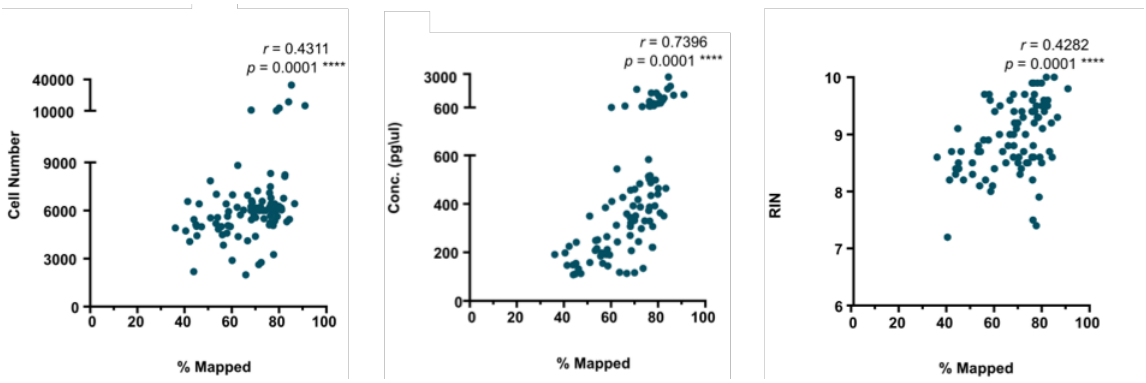


Figure A3.2 RELATIONSHIP BETWEEN THE RNA SEQUENCING LIBRARY ALIGNMENT AND EARLY SAMPLE/LIBRARY METRICS *(n_{sample} = 92). Relationship was quantified by Spearman's correlation. Each dot represents a sample. r - Spearman's rho; p - p-value; RIN - RNA integrity number; Conc. – extracted RNA concentration.*

Despite the increase in read alignment, proportions of genomic origin stayed the same (Figure A3.3).

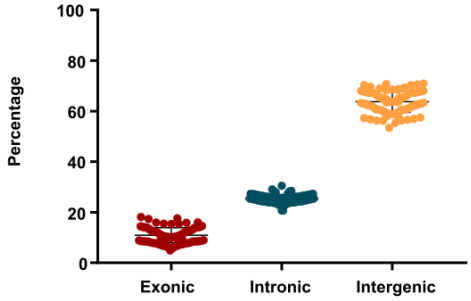


Figure A3.3 GENOMIC ORIGIN OF ALIGNED READS *(n_{sample} = 92). Each dot represents a sample. Error bar represents the mean and standard deviation.*

Together, these data confirmed that the alignment accuracy is determined by the extracted RNA quantity.

Pre-Differential Expression Quality Control

Before differential expression analysis, reproducibility between samples was evaluated. The underlying assumption behind differential expression is that genes from the same sample group have similar expression pattern and quantity.

To evaluate read count distribution, density plots and box plots from normalized counts was generated. No obvious sample outliers were detected (data not shown). Additionally, plots highlighted that most of genes have very low counts (median = 32.7).

DESeq2 intrinsic functions were used for count normalization, variance stabilization and finaly PCA plot and hierarchical clustering generation. PC3/PC4 graphs were produced by manually altering the code used for PC1/PC2 construction. Both the total and protein coding data sets were clustered based on the Euclidean distance. For easy visual assessment of sample associated factor influence on their variance, plots are colour and symbol coded. In addition to colour, sample similarity is highlighted by the dendrograms on top and on left-hand side of heatmap.

In all instances observed clustering was driven by a combination of sample anatomical origin and cell type. When sample relationship was evaluated based on all genes, one of the Blood $CD4^+$ T_{EM} samples was mixed with LPL $CD4^+$ T_{EM}, it was not the case when protein coding genes were assessed alone(Figure A3.4 A).

PCA plots were constructed using expression data for the top 500 most variable genes, as genes that lack variance between samples will not contribute to sample separation. The principal component 1 (PC1) showed that most (48%) of variance is explained by the cell type, following by the anatomical location (PC2 = 27%), as expected. Interestingly, when all genes were considered, as in clustering by Euclidian distance, it was anatomical location that contributed to greatest sample-to-sample difference, followed by cell type (data not shown). PC3 and PC4 explained 4% and 3% of variance, where PC3 seemed to be cell type driven (Figure A3.4 B).

Importantly, PCA showed that samples from LPL have much higher within-group variability than ones from blood, and so each subpopulation should be analyzed separately, if not, LPL samples would inflate the per-gene dispersion estimates for blood samples.

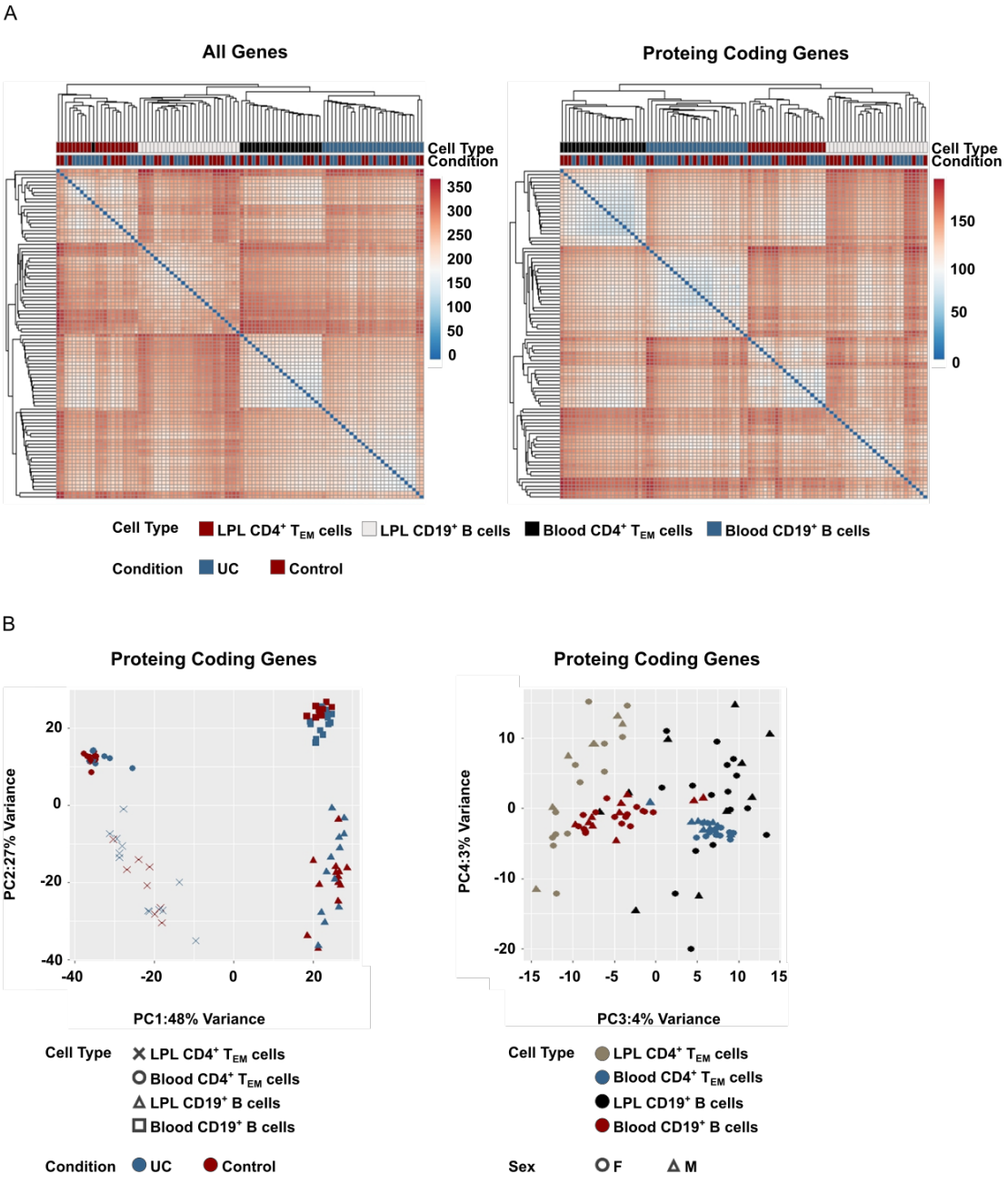


Figure A3.4 EXPLORATORY ANALYSIS (*n samples = 94*). *A. HeatMap representing sample-to-sample relationships at total gene or protein-coding gene level. B. PCA based on top 500 protein coding genes with highest variance. Percentage of variance each principal component constitutes are shown of left-side and bottom of each plot. F – Female; LPL – Lamina Propria; M – Male; PC – Principal component; T_{EM} – T effector memory; UC - Ulcerative colitis patient with and without inflammation in Sigmoid colon.*

After main sample set division into smaller aim specific subsets, PCA was applied to each of individual subsets. To estimate the impact of sample biological and technical variation, PC1/PC2 plots were coloured according to sex, batch (represents RNA sample collection time), medication and disease condition. In addition to PCA, we used *factoextra* R package to generate Scree plots. Scree plot is diagnostic plot showing the percentage of variance explained by each principal component (data not shown). In case of Blood $CD4^+$ T_{EM} and Blood $CD19^+$ B cells, where each PC1/PC2 explains relatively small proportion of variance, PC3 and PC4 must be evaluated too.

PC1 and PC2 (Blood $CD4^+$ T_{EM}) explained 13% and 12% of variance (Figure A3.5 A) and were driven by individual samples; Outlier samples were removed and PCA re-estimated, but again separation was based on individual samples (data not shown) so all samples were kept for further analysis. PC2 (Blood $CD19^+$ B cells) accounted for 11% of variance (Figure A3.5 C) and seemed to separate samples by the RNA sample collection time. However, in addition to sample collection time PC2 slightly correlated with sex, but that could be explained by fact that most of male participants were recruited at second phase of sample collection. In contrast, PC2 (12%) for LPL $CD4^+$ T_{EM} seemed to separate samples by disease state (Figure A3.5 B). The rest of variance could not be explained by any of factors tested (Figure A3.5 A, B, C and D).

A

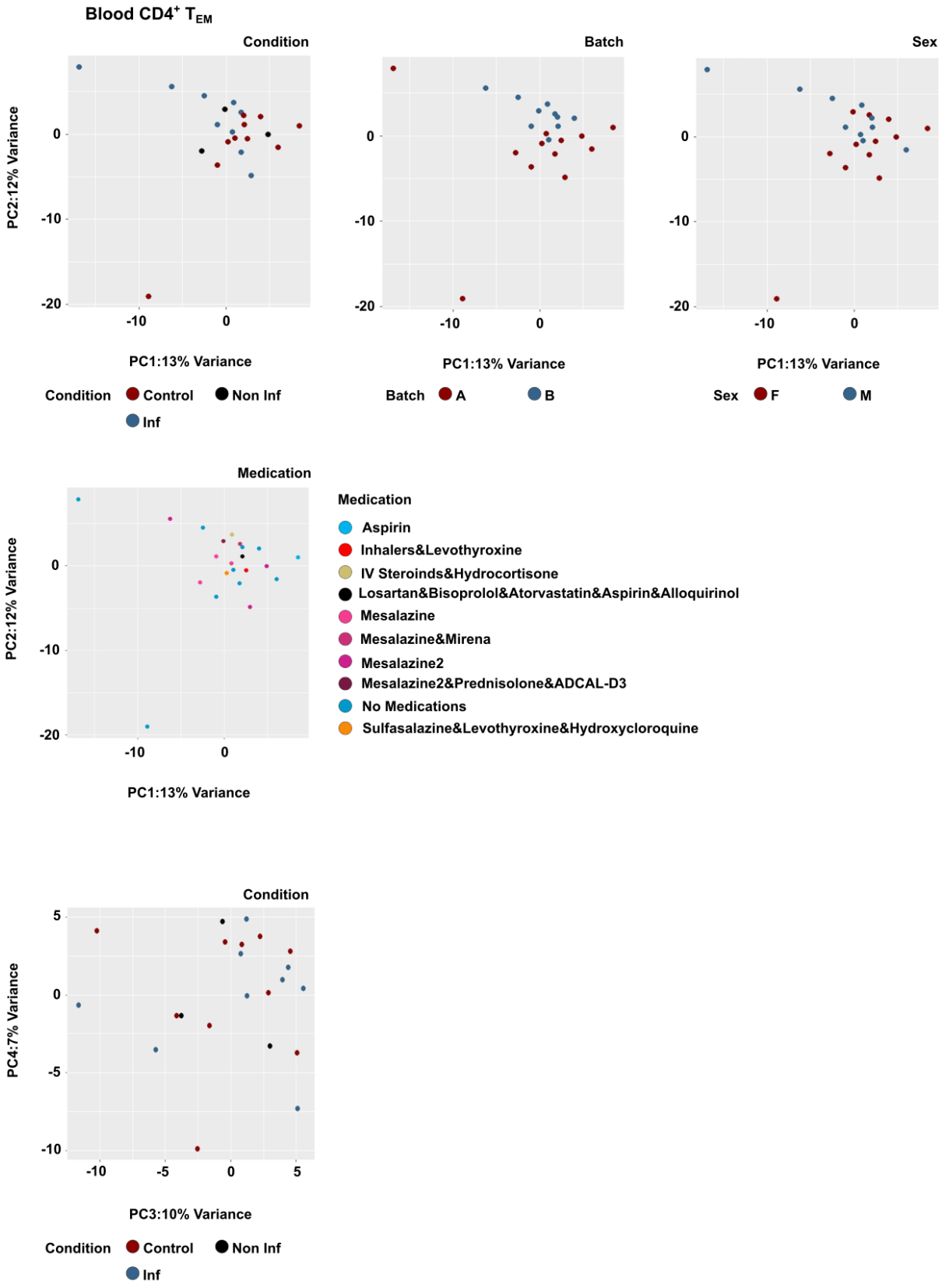


FIGURE CONTINUED IN NEXT PAGE

B

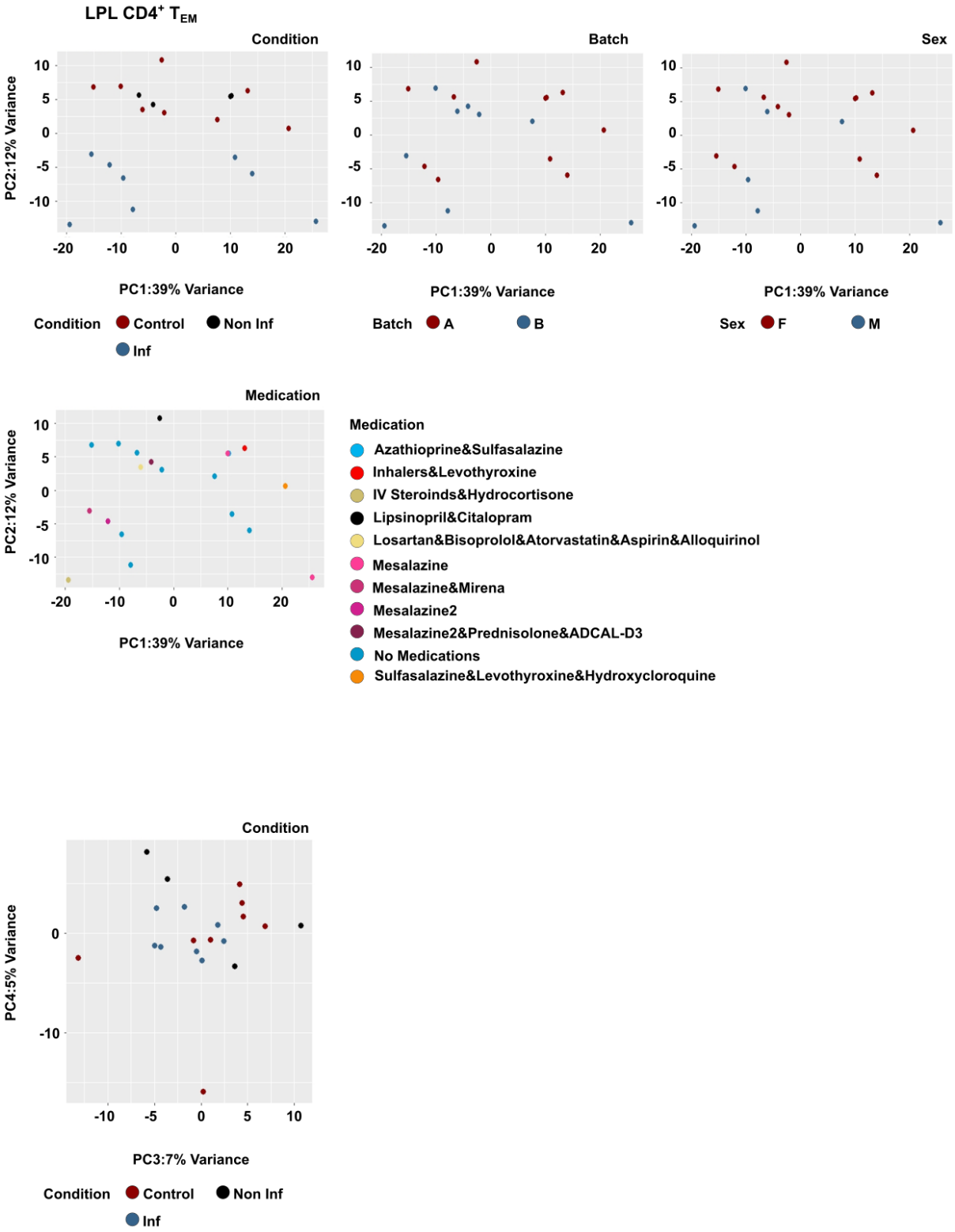


FIGURE CONTINUED IN NEXT PAGE

C

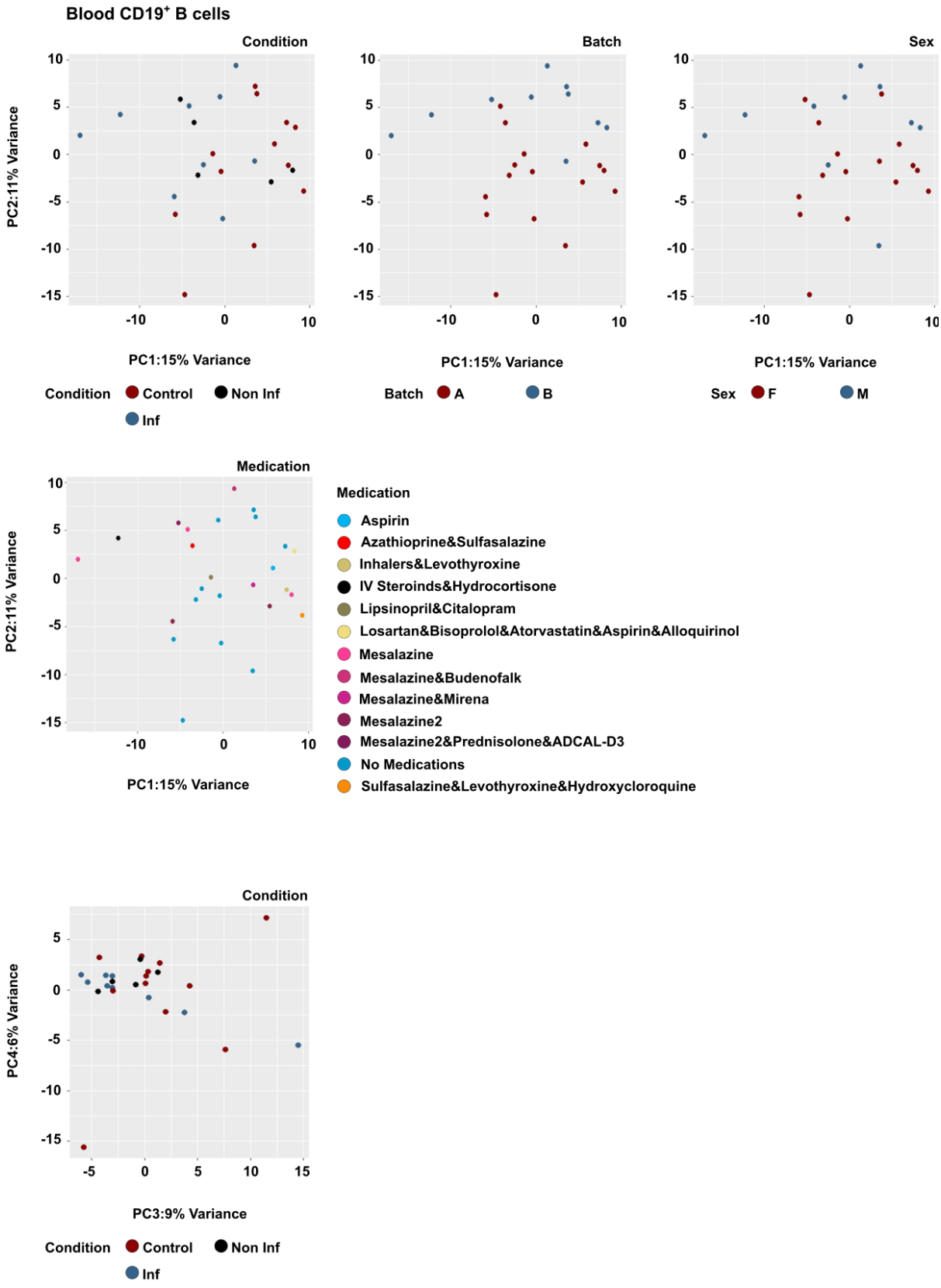


FIGURE CONTINUED IN NEXT PAGE

D

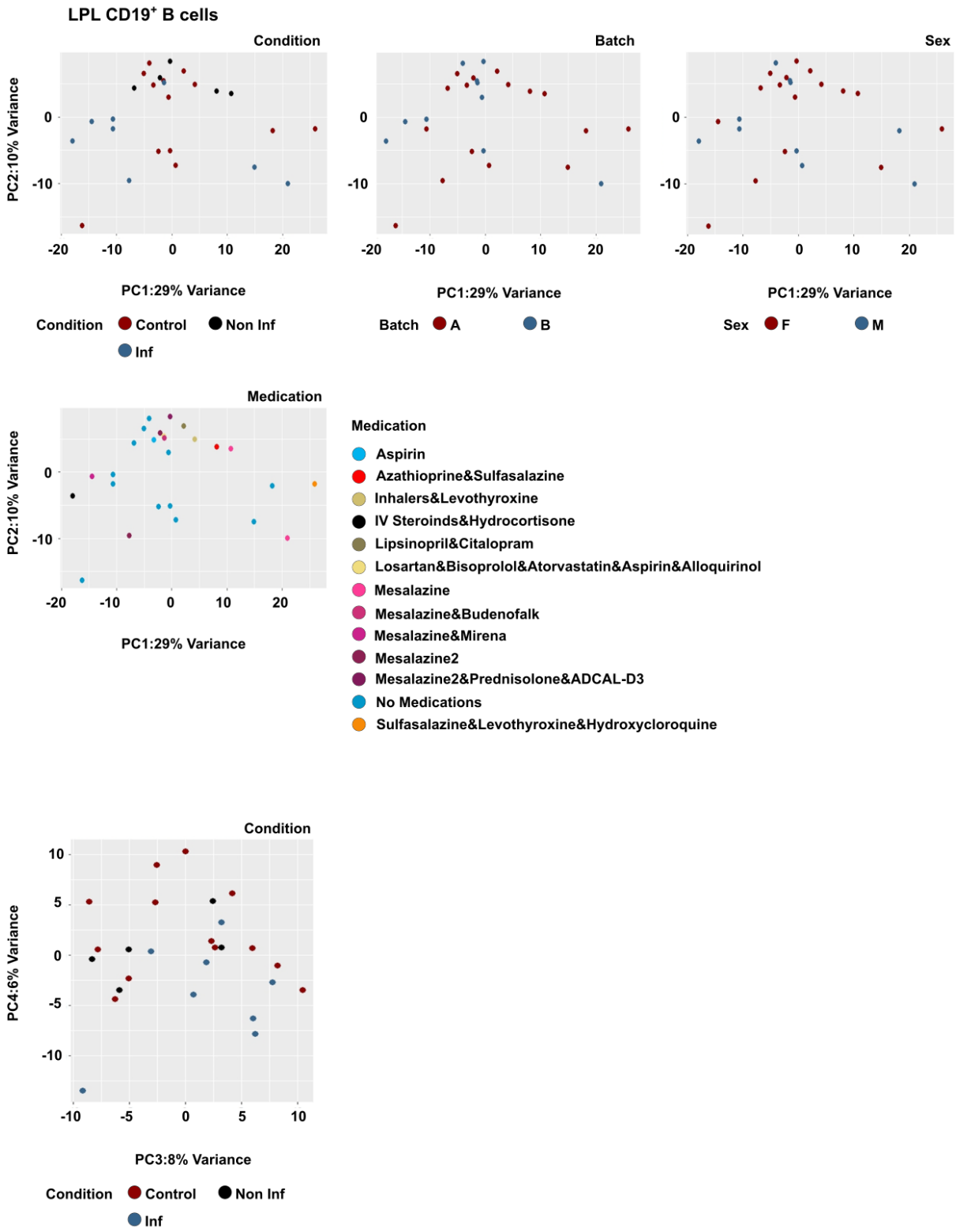


FIGURE CONTINUED IN NEXT PAGE

Figure A3.5 PRINCIPAL COMPONENT ANALYSIS ON A. BLOOD CD4$^+$ T_{EM}, B. BLOOD CD19$^+$ B, C. LPL CD4$^+$ T_{EM} AND D. LPL CD19$^+$ B CELL POPULATIONS. *Batch represents the sample collection and RNA extraction time, where batch A marks samples that were collected 2015-early 2016, but batch B shows samples collected from late 2016-2017. LPL - Lamina propria; T_{EM} - T effector memory; PC - Principal component; F – Female; M – Male.*

When PCA was calculated for samples from the same cell type, but at different anatomical locations, the main variance came from the anatomical location (Figure A3.6), as expected.

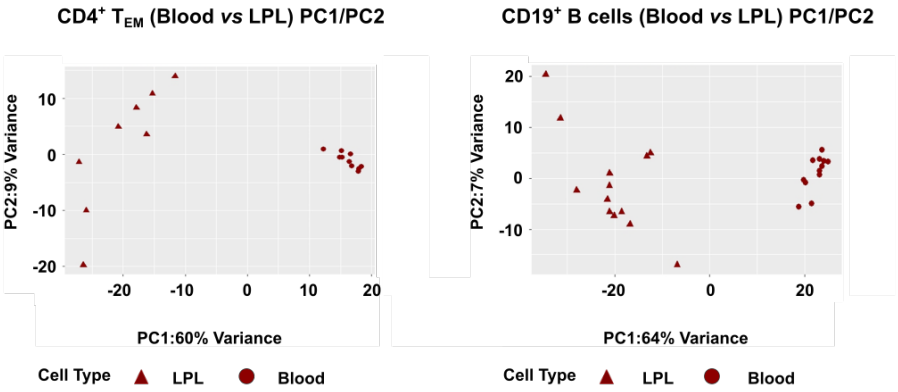


Figure A3.6 GRAPHICAL REPRESENTATION OF PC1/PC2 FOR CD4$^+$ T_{EM} AND CD19$^+$ B CELLS FROM PERIPHERAL BLOOD AND SC LPL. *LPL - Lamina propria; T_{EM} - T effector memory; PC - Principal component.*

For the last step of pre-differential expression quality control, subpopulation purity was evaluated based on their expression of sort markers (CD3, CD4, CD19, CD14, CD62L, CD8a). In initial FACS sort:

- CD45 was used to separate lymphocytes from rest of cells;
- CD3, CD19, CD14 were used to separate all T cells, B cells and MF from CD45$^+$ population;
- CD4 and CD8a was used to separate T cells from CD3$^+$ population.

DESeq2 was used to generate representative plots, where pseudo-count of 0.5 were added to each sample to fit on log axis.

As expected, Blood CD4$^+$ T_{EM} and LPL CD4$^+$ T_{EM} subpopulations expressed more CD4 and CD3 than Blood CD19$^+$ B cells and LPL CD19$^+$ B cells, which was *vice versa* for CD19.

Nonetheless, samples showed marked variation in expression levels of sort markers used for negative selection (Figure A3.7). This could be reflection of low background transcription of genes or simple noise. On the other hand, the sort itself has an error rate, which for samples with lot of cells would make no difference, but for low cell (low count) experiment it might have a rather large impact. Nevertheless, without having a clear understanding of actual reason contributing to heterogeneous expression of negative-sort markers data should be analyses with caution, possibly assuming that there is an underlying contamination.

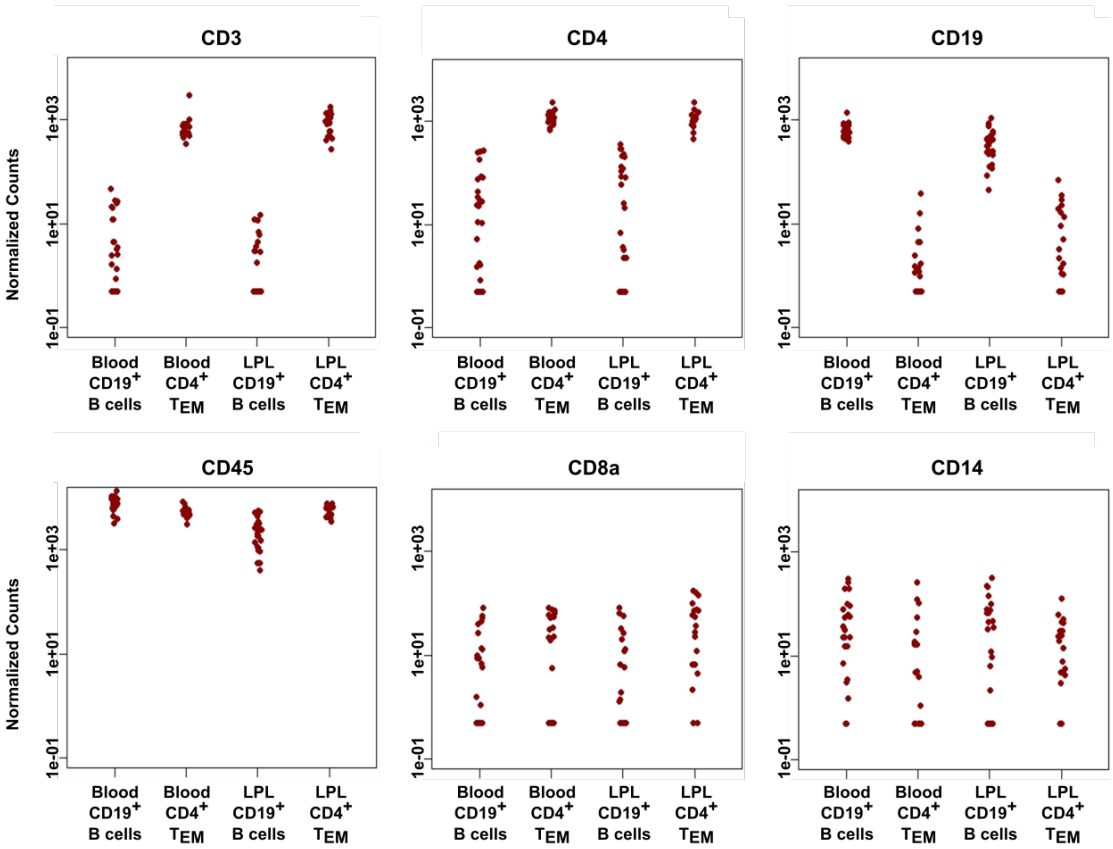


Figure A3.7. EVALUATION OF CELL POPULATION PURITY BASED ON CELL SURFACE MARKER EXPRESSION. *LPL - Lamina propria; T_{EM} - T effector memory; PC - Principal component*

Post-Differential Expression Quality Control

After extensive pre-expression QC the call for differential expression was made. In order to determine if differential expression calculations have resulted in reliable output, we further examined count outliers, independent filtering, p-value and normalized counts for all DEG.

Independent filtering is a metric introduced to minimize the power reduction by multiple testing. Independent filtering works by removing genes with low expression as they suffer from very high Possion noise and are not reliable for estimation of true biological effect (Bourgon, Gentleman and Huber, 2010).

For unknown reasons, most DESeq2 calls were below the independent filtering threshold which meant that function was not performing any filtering. To overcome this and remove genes that had minimal expression, we introduced a manual count-based gene pre-filtering. For full function description please see the Chapter 5 Materuals and Methods section.

DESeq2 function automatically screens for count outliers. The graphical evaluation of gene count outliers showed that none of the DEG had an obvious outlier, neither did any of samples (Data not showed). There is major downfall to outlier identification by DESeq2 - it cannot detect more than one outlier per gene. To further evaluate the quality of DEG calling, normalized counts were plotted for visual inspection for all significant results.

Large proportions of DEG associated with low counts suffered from very high variance (Figure A3.8 A). However, whilst a substantial fraction of DEG associated with low counts were high in variance, the rest displayed good count distribution (Figure A3.8 B).

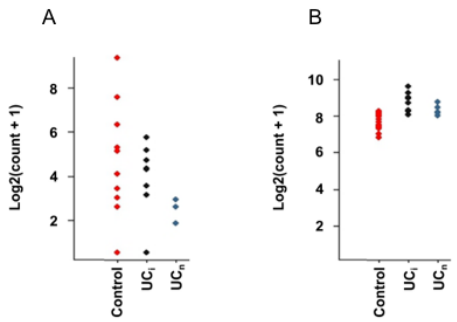


Figure A3.8 REPRESENTATIVE COUNTS DISTRIBUTION FOR A. LOW EXPRESSED GENE AND B. RELATIVELY HIGHER EXPRESSED GENE. *To fit counts of logarithmic axis pseudo-count of 1 was added. UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon; UC(N) - Ulcerative colitis patient with non-inflamed Sigmoid colon.*

Finally, p-value distribution for DEG testing was visualized to determine if the statistical tests employed for significance estimation gave dependable results. p-values should have an uniform distribution between 0 and 1 with a possible peak at 0, representing the p-values under the alternative hypothesis (Klaus *et al.*, 2016). The p-value distribution model in case of little or no significance is named uniform, whereas If there is a significance it is called anti-conservative p-value distribution.

All helathy *vs* disease comparisons had either anti-conservative or uniform p-value distribution indicating the robustness of our approach. However, when comparison was based on anatomical locations, the observed rates of the alternative hypothesis was exceptionally high, suggesting that majority of genes have different expression patterns. To reassess if high number of significant genes is not consequence of test statistics itself, we used *fdrtool* package (Strimmer, 2008) to recalculate the variance of the null-model. By default, test statistics use constant variance *(sd = 1)* for all statistical tests. Yet, if data will have variance different from 0, incorrect estimation will take place. Recalculation showed that, indeed, variance was markedly higher for both $CD19^+$ B cell (Blood *vs* LPL) *(sd = 1.873*) and $CD4^+$ T_{EM} cell (Blood *vs* LPL) (*sd = 1.946*) populations. Calculations was passed back to DESeq2 and p-values corrected according to newly determined variance (Figure A3.9).

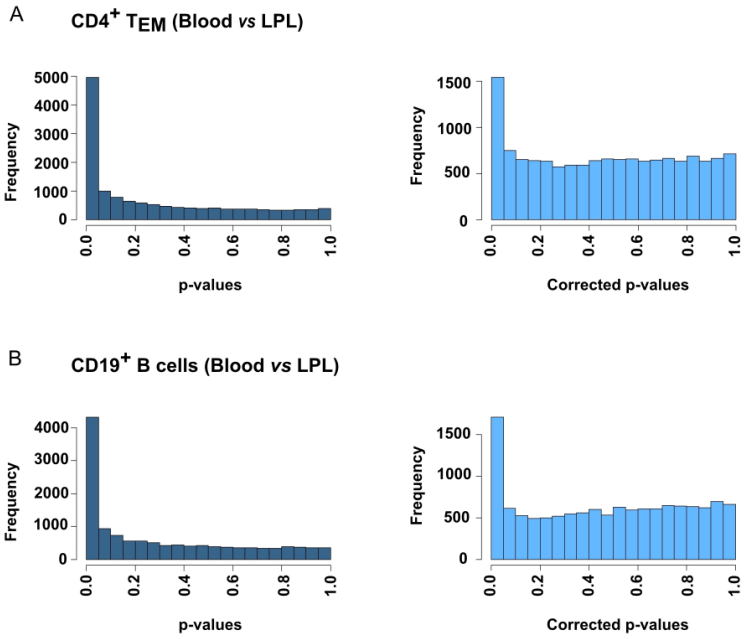


Figure A3.9. P-VALUE VISUALIZATION BEFORE AND AFTER CORRECTION FOR THE STANDARD DEVIATION FOR A. CD4$^+$ T_{EM} (Blood *vs* LPL) AND B. CD19$^+$ B CELLS (Blood *vs* LPL) COMPARISONS. *LPL - Lamina propria; T_{EM} - T effector memory.*

Appendix 4 – Challenges Encountered During RNA Seq Analysis

Poor Alignment To Human Genome

During optimization runs it was discovered that some of the samples showed poor alignment to the human genome. Alignment is one of the first indicators of sample quality and possible contaminations. It was of major importance to identify the factor behind low mapability, as the nature of any confounding factors would determine if samples would meet study expectations and provide an accurate biological picture. Contaminant screening showed that most of the unmapped reads did not belonged to other common bacterial, viral and mammalian genomes. However, correlation with sample associated metrics allowed us to connect poor alignment with RNA poor quality and low yields. We therefore took measures to exclude samples from downstream analysis based upon these parameters.

One critical aspect that impacted library quality related to the library preparation kit used. At the time of designing these experiments, we encountered a paradox. Whilst bulk RNAseq based upon large cell numbers was an established methodology, and whilst single cell methodologies for RNASeq were becoming established, surprisingly little existed for RNASeq library preparation using cell numbers in the range we were exploring. We selected a technology which would allow us to obtain good data quality from clinical samples, known to be limited in nucleic acid amount and heterogeneous in quality. At the time, available kits typically used poly-A selection to separate protein coding mRNA molecules from the rest of RNA species. During mRNA maturation multiple adenosine monophosphates gets added to 3' end of mRNA molecule which is then called poly-A tail (Lodish *et al.*, 2000), which is then targeted for mRNA enrichment (De Klerk, Den Dunnen and 'T Hoen, 2014; Hrdlickova, Toloue and Tian, 2017). mRNA selection is an important step as mRNA constitutes only small fraction of total RNA (Conesa *et al.*, 2016). Thus, without purification it would require very deep sequencing to obtain the same information level as for mRNA alone and lead to high

sequencing costs. Unfortunately, poly-A selection requires large amounts of highest grade RNA as during degradation RNA fragments break and loose their poly-A tails (Conesa *et al.*, 2016). SMARTer Stranded Total RNA Seq Kit - Pico Input Mammalian kit had a distinct advantages in this regard, since it employs bead based rRNA and mitochondrial RNA purification technology which allows kit to be used for suboptimal quality RNA samples, such as those obtained from low starter cell numbers and after some potential degradation during cell sorting and RNA storage. In addition, coupling rRNA removal technology with PCR amplification steps contributed to this kit's unique picomolar RNA input requirements (Takara Bio Inc., 2018).

In our experiments, we showed that in our hands, this kit was not compatible with poor quality RNA samples. To gain further understanding of why we see low quality in our sequencing library, I undertook further communication with Dr Bergamaschi, a post-doctoral fellow in Prof Ken Smith's laboratory and who had routinely used SMARTer Stranded Total RNA Seq Kit - Pico Input Mammalian kit for her sequencing library generation. Dr Bergamaschi did not share our experience, yet, their laboratory used ng of only high-grade RNA (RIN > 7). Indeed, our observation was latter backed up by *Schuierer et al 2017.* They compared poly-A and rRNA removal technology back to back across a wide range of RNA input concentrations coupled with different RNA quality and showed that sample alignment decreased with reduced input amounts independent of RNA quality, and that highly degraded samples resulted in very poor alignment to human genome which was further augmented by a decrease in input concentration. The lower concentration of highly degraded RNA tested on rRNA depletion kit in their study was 10ng (94% more than for our libraries) which produced only 42.78% alignment to human genome (Schuierer *et al.*, 2017).

Unexpected Genomic Origin

In addition to poor alignment, a separate problem we observed was that for mapped reads, there was an unexpected pattern of genomic alignment, with the highest fraction of reads mapping to intergenic part of DNA which should not be represented in RNA. mRNA selection-based studies usually showed high (70% - 90% of mapped reads)

exonic alignemnt and high intergenic mapping acted as first indicator for gDNA contamination. To exclude possible gDNA contamination (Griffith *et al.*, 2015; Conesa *et al.*, 2016), a series of further tests were performed, all of which did not support the presence of significant gDNA contamination in our samples. However, we did not employ any further computational gDNA evaluation methods, but visualization of intergenic reads could be a good experiment to start with. If gDNA was responsible for intergenic reads uniform read distribution would be expected (Andrews, 2016b). Another, good in silico experiment to resolve the high intergenic alignment would be looking for intergenic regions unusually high in reads, which would be an expected result if any of random primes showed binding preference.

However, we were aware that rRNA depletion technology might change the read distribution as it depletes only rRNA and mitochondrial RNA but preserves the rest of RNA species. Around the time of our observations, available data from Illumina appeared to supported the effectiveness of their rRNA depletion technology that we employed. However, since this time, several publications have compared different RNA Seq technologies and have clearly shown that sequencing libraries generated by rRNA depletion have lower exonic read alignemnt. With that being said, there currently are no consensus of what the expected alignment pattern for rRNA depleted libraries is. For example, a recent study by *Herbert et al 2018* compared seven different rRNA depletion techniques. They used the same commercially available RNA reference sample (The Universal Human Reference RNA (Agilent)) and rRNA depletion method as *Schuierer et al 2017*. Nonetheless, they were not able to reproduce the same genomic alignment as the other group. The final two messages conveyed by *Herbert et al* were that the same genes were detected significantly different by different kits and that performance of the same kit in different RNA seq facilities were markedly different (Herbert *et al.*, 2018).

Interestingly, samples of different origins (Schuierer *et al.*, 2017; Zhao *et al.*, 2018) and preservation techniques (Zhao *et al.*, 2014) have high variance in their exonic : intergenic : intronic read ratios. *Zhao et al 2018* showed that in order for rRNA depleted blood and colon samples to reach the same protein coding gene coverage as produced

by poly-A selection 220% and 50% more reads would be required, making rRNA technology very cost ineffective if it is protein coding genes which are of interest (Zhao *et al.*, 2018).

In conclusion, based upon our experiments and those of others, we believe that currently none of the kits for sequencing library construction from bulk of RNA would yield good quality sequencing libraries when small clinical samples with only pg of suboptimal quality available. Instead, if our experiments were to be repeated, methodologies based upon single cell sequencing should be considered. Decreasing costs and increased throughput rates makes it this an attractive means to study the transcriptional profiles in cell populations present in clinical samples. However, the downfall for single cell technology is the limited number of transcripts one gets when compared to bulk sequencing. Indeed, during the review of the present manuscript, we became aware of recent related work by Smillie et al. ('Rewiring of the cellular and inter-cellular landscape of the human colon during ulcerative colitis', submitted for publication). These authors proceeded to generate a single-cell expression atlas from the colon to capture the transcriptional landscape of healthy and UC patients. Together their work showed that genes with high probability of being causal share very cell subset- and linage- specific expression patterns, which change upon inflammation (Smillie *et al.*, 2018). This observation further compliments our work, where we proceed to investigate the transcriptional changes and, later, epigenetic changes in cell type specific manner.

Appendix 5 –Power Calculation For RNA seq Experiments

Next, we proceeded to estimate what sample size is needed to correctly reject the null hypothesis that there is no difference in gene expression between two groups. We used *RNASeqSampleSize* package developed by *Zhao et al., 2018*. It should be noted, that due to RNA seq data complexity (they represent hundreds of genes with different expression values, distribution and dispersions) there is no standard method for Power Calculation. In addition, user must provide with descriptive values for prognostic genes (genes that will be differentially expressed). Therefore, to have the best possibility to accurately capture sample size needed, we used two different functions from *RNASeqSampleSize* package using our own data and publicly available large colorectal cancer RNA seq data set.

First, we created power curves using dispersion values from our data. In short, we asked what will be the power to accurately reject null hypothesis if:

1. Prognostic genes will have average expression of either 100 counts (lambda0 = 100) or 250 counts (lambda0 = 250).

2. Minimal fold change associated with prognostic gene is 2, 3 or 4 (rho = 2, 3 or 4)

Sample numbers used for RNA seq data analysis is showed by white interrupted lines. When sample numbers in each group was not equal, such as 12 vs 9, we used the lower value. We reasoned that it will give more realistic picture (due to higher chance to have estimated population mean deviated from real population mean).

Calculations showed that Control vs UC_n comparison were too low in sample size to accurately reject null hypothesis even if prognostic gene has 4-fold change between conditions compared. Whereas Control vs UC_i comparison would be able to capture

very prominent prognostic genes that has high median cell counts and fold change associated with them (Figure A5.1 and A5.2).

As power depends on dispersion we wanted to use large data set with many samples to give us the best chance to accurately capture intra-group variance in expression. We asked how many samples would be needed to have 80% possibility that we have accurately rejected null hypothesis in data set with minimal average gene expression of 10 counts, with total number of expressed genes ~8K.

Calculations showed that if prognostic gene has minimum difference of 2 folds, 3 folds or 4 folds, 48, 20 and 12 samples are needed, respectively. Altogether agreeing of what we saw using our own data.

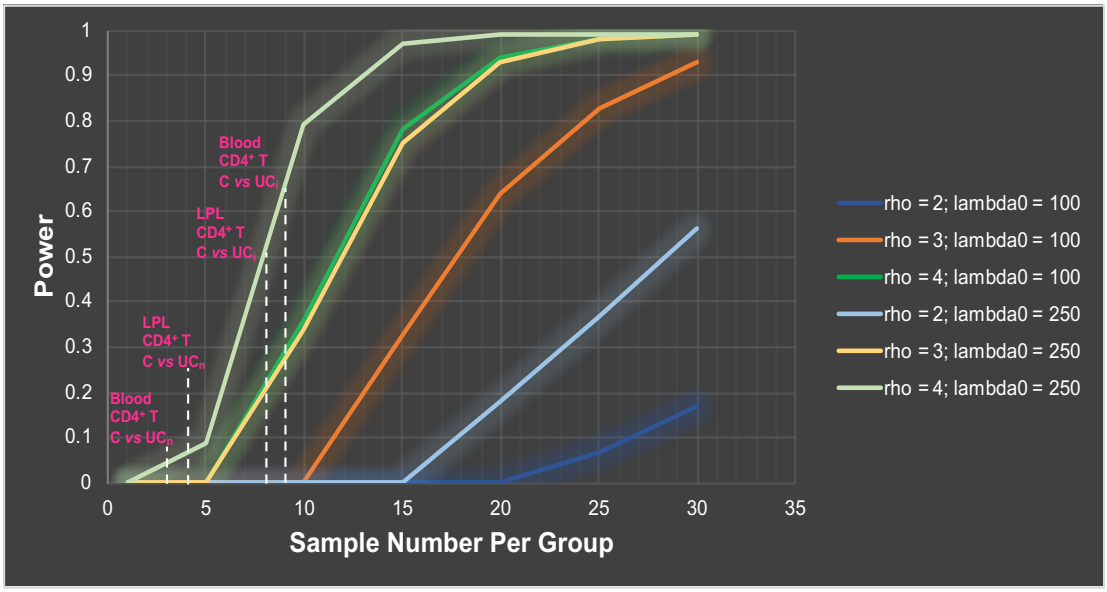


Figure A5.1 POWER CURVE FOR CD4⁺ T CELL DATA. x axis represents sample numbers, whereas y axis shows the associated Power. Both Blood and LPL CD4⁺ T cell data had the same dispersion value of 0.6 and 0.4 for genes with average counts of 100 and 250, respectively. Rho shows the minimum expected fold change and lambda0 represents the average counts for prognostic gene. White lines represent sample numbers in our analysis.

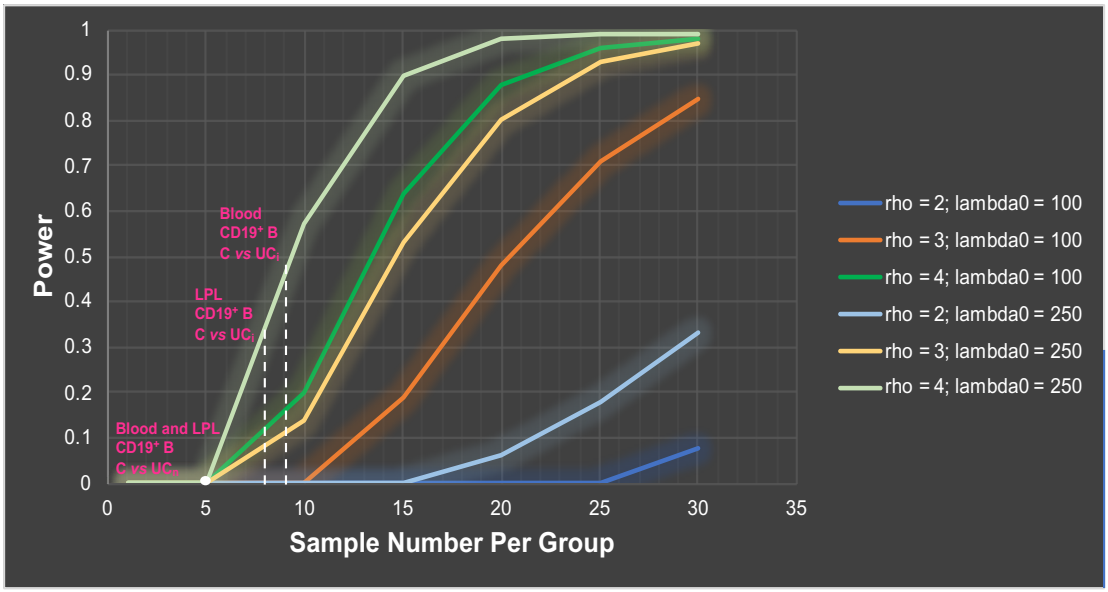


Figure A5.2 POWER CURVE FOR CD19$^+$ B CELL DATA. x axis represents sample numbers, whereas y axis shows the associated Power. Both Blood and LPL CD19$^+$ B cell data had the same dispersion value of 0.7 and 0.5 for genes with average counts of 100 and 250, respectively. Rho shows the minimum expected fold change and lambda0 represents the average counts for prognostic gene. White lines represent sample numbers in our analysis.

In summary, we showed with the current sample numbers we have up to 0%-8% probability to accurately reject the null hypothesis when Control is compared to the UC_n populations, when prognostic gene has average counts of 250 and fold change of 4. However, in LPL CD19$^+$ B cells (C *vs* UCn) we have identified gene *HIST1H2AJ* that has median counts of ~1411.04 and Log2fold change of -3.65 and has 99% probability to be an accurate estimation.

When Control samples are compared to UC_i populations, we have 50% LPL CD4$^+$ T cells (C *vs* UCi), 65% Blood CD4$^+$ T cells (C *vs* UCi), 34% LPL CD19$^+$ B cells (C *vs* UCi) and 49% Blood CD19$^+$ B cells (C *vs* UCi) chance of correctly rejecting null hypothesis if prognostic gene has 4 fold (Log2Fold of 2) difference and average read counts of 250. Altogether, we wanted to highlight that this calculation is highly dependent of the parameters we associate with prognostic genes. Hereby, we conclude that our Control vs UC_i data sets could be used to identify prognostic genes with high average expression and/or fold change. However, the further validation experiments are crucial before making any biological conclusions.

Appendix 6 – Extended ATAC Seq Data QC

Initial ATAC Seq analysis was performed by Dr. J. Gutierrez-Achurry (postdoctoral research associate in Dr Carl Anderson's group, Wellcome Trust Sanger Institute). Further analysis was carried out by the author. ATAC Seq Data quality was determined at various stage of analysis and has been main findings has been summarized below.

Raw Sequencing Read Alignment And Assessment Of ATAC Library Quality

Substantial variation in sample sequencing depth was observed with total raw PE read count ranging from 11.2M to 225.6 M reads/sample with median of 49.6M reads/sample (Figure A6.1). Three samples particularly separated from the rest, these all came from a single sequencing lane which had been loaded with only these 3 samples, resulting in less competition for binding.

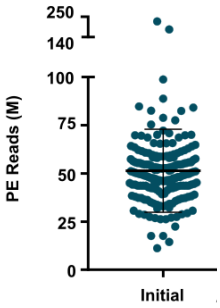


Figure A6.1 ATAC LIBRARY SEQUENCING DEPTH (n $_{sample}$ = 183)**.** *Error bar represents the mean and standard deviation. PE - Paired end; M – Million.*

Following initial clean-up which included adapter trimming and short fragment removal, raw reads were mapped to human genome (version 38) by use of BWA aligner (Li and Durbin, 2009). The vast majority (89%) of samples had a mapping score higher than 90% (Figure A6.2), signifying overall sequencing accuracy and absence of other contaminating genomes.

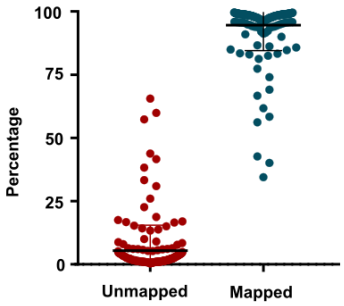


Figure A6.2 ATAC SEQUENICNG LIBRARY ALIGNMENT TO HUMAN GENOME EXPRESSED AS PERCENTAGE OF TOTAL RAW READS (n_{sample} = 183)**.** *Error bar represents the mean and standard deviation.*

The human mitochondrial genome is a well-recognized contaminant of first generation ATAC Seq libraries and this requires removal from datasets. To estimate how much of the total read counts were taken up by reads mapping to the mitochondrial genome, Samtools (Li *et al.*, 2009) command *indexstats* was run by Medimmune Bioinformatician. The output file was processed by author and numbers expressed as percentage of total library size (Figure A6.3).

Mitochondrial read counts in samples fluctuated, with some samples having only 0.11 % of total read counts taken up, but others as high as 73.20%. High presence of mitochondrial reads meant that reads from nuclear origin (of interest) will have reduced sequencing depth in these samples and, possibly, will lead to decreased power to detect the difference in chromatin conformation (hereafter referred to as differentially accessible regions (DAR)) between the control and UC.

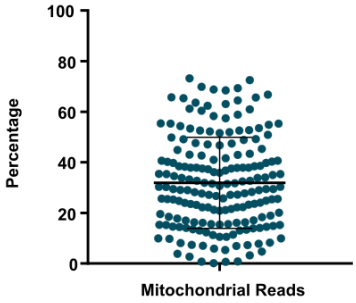


Figure A6.3 PERCENTAGE OF TOTAL PROCESSED READS THAT MAPPED ON MITOCHONDRIAL GENOME (n $_{sample}$ = 183)**.** *Error bar represents the mean and standard deviation.*

To obtain an impression of final library depth after all filtering, total read counts from each purified read file was visualized (Figure A6.4). Initially we set a goal for 40M PE reads/sample which should result in 20M SE reads/sample, yet we only achieved a mean of 7.5M SE reads/sample, thus more than half of libraries had lower counts that had been previously determined as optimal.

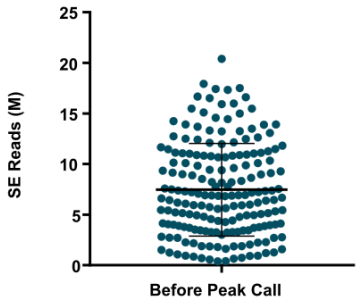


Figure A6.4 READ COUNT PER SAMPLE POST FILTERING THAT WAS USED FOR PEAK CALLING (n $_{sample}$ = 183)**.** *Error bar represents the mean and standard deviation. SE - Single End; M - Million.*

For further analysis samples were selected based on their Fraction of reads in peaks (FRiP) and total read count. FRiP was calculated by dividing the number of reads falling in peak regions with total mapped read count and has been established as a quality metric for Chip seq data interpretation (Landt *et al.*, 2012). The recommended cut off threshold used in ATAC Seq studies ranges from ≤10% to ≤15% of FRiP. In this study we selected a less conservative cut off of ≤5%, due to the large number of samples falling below the 10% margin (Figure A6.5).

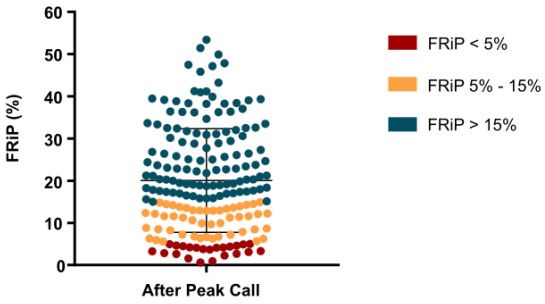


Figure A6.5 PERCENTAGE OF FRACTION OF READS IN PEAKS (n $_{sample}$ = 183)**.** *Red dots represent samples with FRiP less than 5%, Orange dots shows samples with FRiP higher than 5% and lower than 15%, whereas blue dots represent samples with FRiP higher than 15%. Error bar represents the mean and standard deviation. FRiP- Fraction of Reads in Peaks.*

In addition to FRiP, reads count for each sample in the group it belonged to (e.g. Blood $CD4^+$ T_{EM}) were compared to median read count for the same group. The individual read count per sample was expressed as percentage of median read count, and samples that fell below 10% on this assessment were removed. Applying these filtering criteria resulted in a total of 25 samples (from the original 183) being excluded from further study; interestingly 9 of these 25 samples came from the same donor, indicating on possible error in library preparation stage.

Pre-Differential Accessibility Quality Control

Before differential accessibility calculation can commence, we performed visual data exploration to ensure that samples are indeed representing the expected phenotype and there are no other factors that would drive sample separation.

Unfortunately, we were not able to used DESeq2 packadge inherit exploratory analysis, as at the time of analysis DiffBind packadge did not provide this functionality. It is preferable to use the same normalized counts for significance estimation and initial visualization as each normalization metric influences sample to sample relations. Instead, pre-filtered raw read counts were normalized for sequencing depth and length of each peak by RPKM and log2 transformed to minimize the influence on test statistics of peaks with the highest counts. Samples clustering was assessed by Spearman correlation and by PCA.

Sample correlation revealed that the main variable driving sample separation was cell type, where Blood $CD14^+$ MF, LPL and Blood $CD19^+$ B cells and $CD326^+$ Epithelial cells made distinct groups with most of $CD19^+$ B cell samples further separating based on their anatomical region (Figure A6.6). Yet, T_{EM} clustering was much more heterogeneous with both $CD8^+$ and $CD4^+$ cells from all three regions combining in smaller patches. One of the plausible explanation for T_{EM} patchiness would be that both- anatomical location and cell type had similar weights on sample clustering. Interestingly, a small subset of the T_{EM} samples were closer to epithelium. We were not sure what factors were behind this separation, but contamination seemed less likely as some of the separated T_{EM} samples were of peripheral blood origin. Similar patterns were observed when clustering was based on top 500 and top 5000 peaks with highest standard deviation (data not shown).

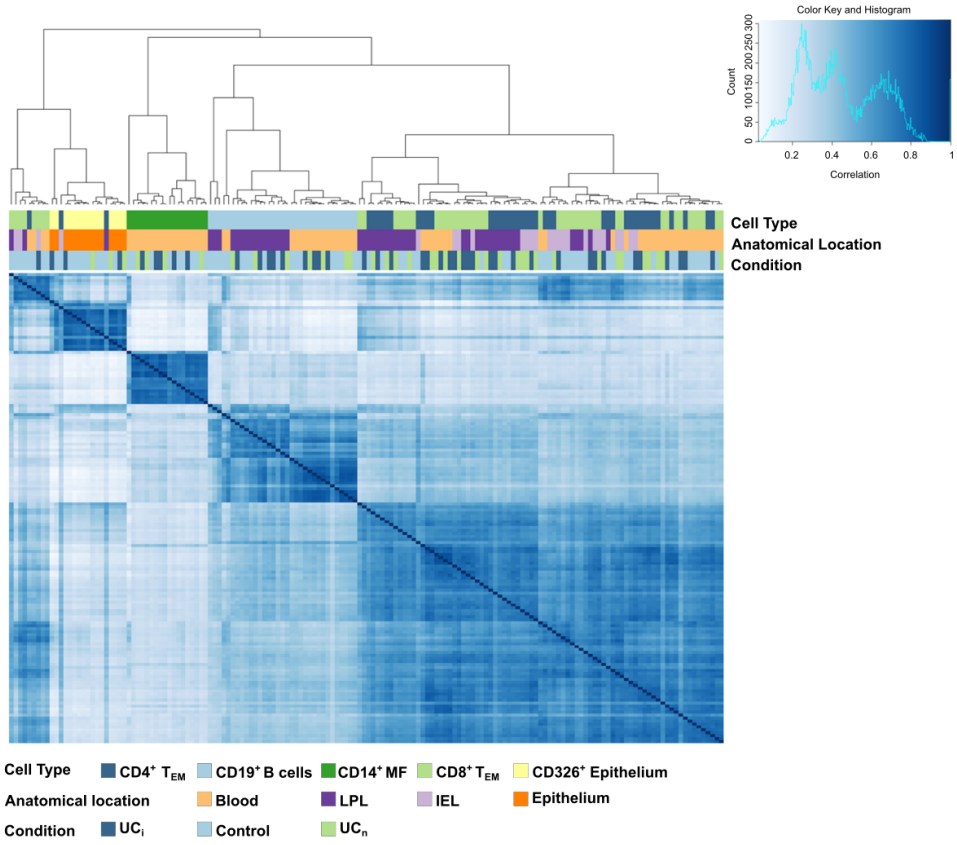


Figure A6.6 HEATMAP REPRESENTING SAMPLE-TO-SAMPLE RELATIONSHIPS *($n_{samples} = 158$). In colour-key 1 (the same sample) is represented by dark blue colour, which slowly transitions into white with samples becoming more different to each other. Samples from UC_i, UC_n and Control patients were colour-coded dark blue, green and light blue, respectively. Light orange, dark violet, light violet and dark orange were assigned to Blood, LPL, IEL and Epithelium. Finally, dark blue, light blue, dark green, light green and yellow shows $CD4^+$ T_{EM}, $CD19^+$ B cells, $CD14^+$ MF, $CD8^+$ T_{EM} and Epithelium. LPL – Lamina Propria; IEL – Intraepithelial Lymphocytes; UC(I) - Ulcerative colitis patient with inflamed Sigmoid colon; UC(N) - Ulcerative colitis patient with non-inflamed Sigmoid colon; T_{EM} – T effector memory; MF - Macrophages and Monocytes.*

To investigate how known technical and biological factors such as disease state, secondary condition, sex, medication, total PCR cycles and Tn5 enzyme volume used for library construction and sorter type, might influence sample clustering, PCA for individual *binding affinity matrices* were calculated and samples in PC1/PC2 plots color-coded based on feature they represented.

In all 10 models, none of the factors tested showed any influence on PC1/PC2 (data not shown).

Post-Differential Accessibility Quality Control

In the same manner as for RNA seq data analyis, we perfomed vigorous assessment of test statistics.

Data passed all but the p-value distribution QC metrics. Instead of anti-conservative or uniform p-value distribution these plots had conservative (“hill-shaped”) distribution with more p-values at 1 than at 0. The odd p-value distribution is first indicator that test statistics has failed and results are not reliable for any furter interpretation. Representative plots are showed in Figure A6.7.

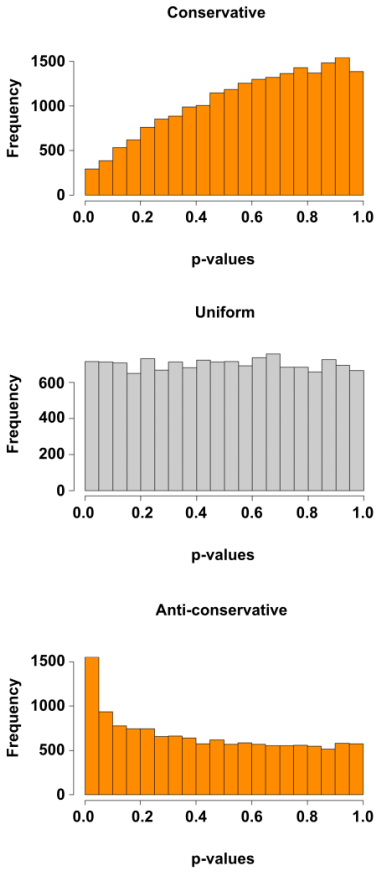


Figure A6.7 P-VALUE HISTOGRAMS ILLUSTRATING CONSERVATIVE (HILL-SHAPED), UNIFORM AND ANTI-CONSERVATIVE DISTRIBUTIONS.

Determining The Possible Causes Behind Failure For Statistical Testing

We performed rigorous search of any previously reported instances when the DESeq2 test statistics had failed and returned hill-shaped p-value distribution. The 3 most common reasons we found were:

1) If variance from the null distribution is too high (Klaus *et al.*, 2016).

2) Unidentified batch effect (Bergenstrahle, 2017).

3) Overwhelming prescience of genes with low counts (in our case peaks) (Bergenstrahle, 2017).

To permitted more flexibility for data assessment, we transformed all individual *DiffBind class objects* into the DESeq2 specific objects. Next, we proceeded to investigate if any of 3 proposed reasons could explain failure of statistical test seen.

First, we used the *fdrtool* package to re-evaluated the null variance for all comparisons. However, thought the most of calls had variance different than 1, the newly estimated variance did not rescue the test statistics.

Next, to test if any of known technical or biological factors influence sample relation, PCA was re-performed. This time PCA was calculated by DESeq2 internal function. In addition, we decided to use *variance stabilizing transformation* instead of log2 transformation. *Variance stabilizing transformation* accounts for both - high and low count impact on variance.

Despite data pre-filtering and improvement in normalization and variance stabilization, it was still very hard to dissect if and how much of any factors tested had an influence on sample variance. However, the message from newly generated PC plots were consistent with all previous results (Data not shown).

Following the PCA assessment, the raw and normalized counts for each sample were visualized by boxplots and density plots. This showed that some of the samples were so low on counts that even after normalization they failed to align with other samples.

Observed low count dominance was not unexpected, as most samples were left with small library sizes after initial raw count clean-up had taken place. However, further sample exclusion would lead to reduced power to detect DAR. Nevertheless, keeping outlier samples which resist normalization appeared to have a more negative impact on test statistics, so these outlier samples were removed after manual inspection. In addition, stringent count pre-filtering was introduced, stating that each sample in comparison must have 10 or more counts for a given region to be retained.

Finally, the call for the differential accessibility was repeated and test statistics evaluated. Thought, now only 5 of 20 tests showed a hill-shaped p-value curves the number of differentially accessible regions identified between Blood $CD19^{+}$ B cells $_{(C\ vs\ UCn)}$ and $_{(C\ vs\ UCi)}$ were very high making us question if some of the calculations have truly worked.

In summary, majority of samples in this experiment are under sequenced and below our target of 40M reads/sample. Data set can be used to identify peaks which are present/absent in each phenotype, yet the differential expression analysis must be looked with caution, particularly as our test statistics either failed or showed suspiciously high number of differentially accessible regions. In addition, the inter-group variance is very high suggesting that very small sample numbers we have are insufficient to correctly represents estimated population mean, thus has little or no power to infer population based conclusions.

Appendix 7 – Count Normalization Methods And Their Application In ATAC Seq Analysis

At time of ATAC seq data analyis there were no standard methodologies established for ATAC seq data normalization. The normalizetion method that the DESeq2 model uses accounts for both - differences in sequencing depth and sample composition (Love, Huber and Anders, 2014). In most published ATAC Seq studies authors normalize for depth alone, and do not seek to correct for sample compositional bias. In order to evaluate if DESeq2 inherit data normalization metrics would be suitable for ATAC seq data, we proceeded to performed a series of data symulation experiments.

First we looked to assess if the read count in peaks could be used as representation of the toatal sequencing depth. The total library size (i.e. reads used for peak calling) was correlated with the total read count in peaks for each library, and strong, positive correlation (*r = 0.74, p = 0.0001*) observed (Figure A7.1). This indicated that the metric of total reads in peaks could be used to adjust for difference in library size.

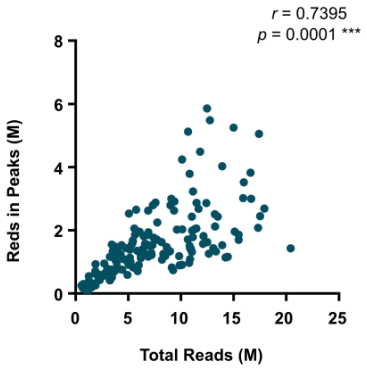


Figure A7.1 SPERAMAN CORRELATION LOOKING AT THE RELATIONSHIP BETWEEN THE TOTAL LIBRARY SIZE AND READS IN PEAKS (*$n_{samples} = 158$*). *Relationship was quantified by Spearman's correlation. Each dot represents a sample. r - Spearman's rho; p - p-value; M – Million.*

Next, to investigate if DESeq2 normalization metrics are compatible with ATAC Seq data, both median ratio and scaling to full library size were compared head to head. For data symulation purpose, a new counts matrix from Blood Control $CD4^+$ T_{EM} and $CD19^+$ B cells were generated. We worked under assumption that genes under active transcription should have their TSS open (Boyle *et al.*, 2008). Since transcriptional differences between healthy blood T cell and B cells are well established, looking at differentially accessible peaks and linking these with the TSS of nearby genes, then comparing these to known T and B cell DEG, represents a strategy for differential accessibility pipeline validation. In this study TSS region was set to ± 1000kb upstream and downstream from the actual TSS (according to the definitions used in *de la Torre-Ubieta et al., 2018*).

A total of 15988 and 15284 DAR between blood control $CD4^+$ T_{EM} and $CD19^+$ B cells were identified when read counts were normalized by median ratio or scaling method, respectively. 82.85% of all DAR peaks were identified by both methods (Figure A7.2).

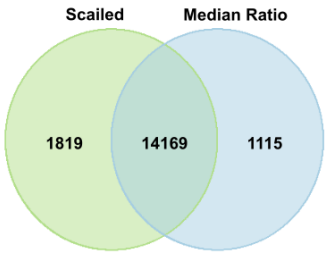


Figure A7.2 VENN DIAGRAM REPRESENTING THE NUMBER OF DIFFERENTIALLY ACCESSIBLE REGIONS IDENTIFIED FROM THE SAME DATA SET NORMALIZED BY EITHER MEDIAN RATIO METHOD OR SCALED TO FULL LIBRARY SIZE.

To further investigate if the peaks identified were consistent with expected biological differences, all differentially open regions were annotated for their genomic origin and assigned to a gene (by proximity to nearest TSS). Associated genes were then analysed using KEGG pathway enrichment analysis (Zhao, Guo and Shyr, 2019).

DA regions were enriched for B cell receptor signalling pathways and T helper 1 and T helper 2 differentiation (Figure A7.3), providing reassurance that DESeq2 internal

normalization metrics can be used for current ATAC Seq data set and that identified peaks represent true biology.

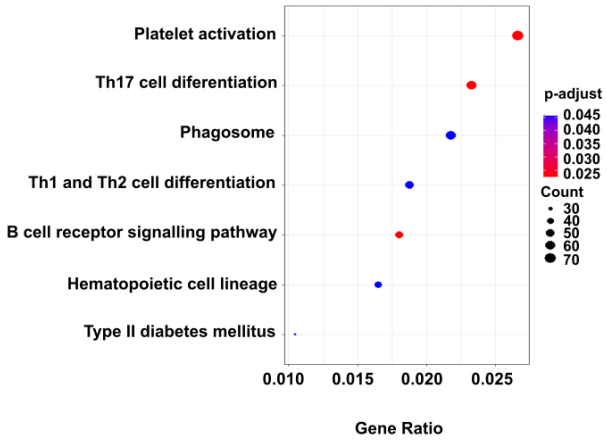


Figure A7.3 KEGG PATHWAY ANALYSIS OF CHROMATIN REGIONS IDENTIFIED AS DA BY MEDIAN RATIO METHOD. *T_{EM} - T effector memory cell;*

Appendix 8 – Challenges Encountered During ATAS Seq Analysis

High Mitochondrial DNA Contamination

DA estimation was far from straightforward. First, substantial percentage of reads (median = 30.2%) were lost due mitochondrial contamination, which is an expected problem for all first generation ATAC Seq protocols (Gaspar, 2019). Mitochondrial DNA is nucleosome-free (Alexeyev *et al.*, 2013) which makes it an ideal substrate for transposase cleavage. Since recognition that large proportion of sequencing reads are lost due mitochondrial contamination various mitochondrial depletion techniques has been evaluated (*Wu et al., 2016*; *Gu et al., 2016*; *Corces et al., 2017;* Montefiori *et al.*, 2017).

Assessment of Omni-ATAC Seq protocol allowed us to propose a reason behind the very heterogeneous amounts of mitochondrial DNA seen in our libraries. Even though our nuclei extraction step was adopted from DNase I hypersensitivity assay (John *et al.*, 2013) the detergent and order of tagmentation steps stayed the same as in the ATAC Seq protocol. Both protocols use IGEPAL - a nonionic, non-denaturing detergent, and immediately proceed with transposition reaction. We believe that incomplete lysate removal from the nucleic fraction has resulted in mitochondrial contamination. Due to low cell numbers the nuclei pellet was not visible after initial centrifugation following lysis. Thus, in order to avoid removal of nucleic material some lysate was left at the bottom of tube. Additionally, the lysis reaction was carried out in 1.5ml Eppendorf tubes, further complicating the identification of any nuclei pellet. Therefore, the inconsistency in lysate removal would explain the highly heterogeneous amounts of mitochondrial reads seen.

Increased Background Signal

Next, 25 samples were effectively lost due low signal to background ratios. Hoverer, the actual number of samples with suboptimal quality (FRiP < 15%) was 71 which

would correspond to more than 1/3 of all samples sequenced. Currently, we do not know the reason behind the low signal to background ratio. However, a literature search has allowed us to identify 3 known factors which influence tagmentation. They are chromatin conformation, underlining DNA sequence (Madrigal, 2015)and Tn5 transposase concentration (Buenrostro *et al.*, 2015).

If ATAC Seq libraries have been constructed correctly, around half of post-cleanup fragments should be shorter than 100bp and the rest of the fragments should show a prominent downward laddering pattern (Buenrostro *et al.*, 2013; Ou *et al.*, 2018). The small fragments (<100bp) are representative of nucleosome free regions, whereas the downwards facing ladder corresponds to nucleosomes tightly wrapped in chromatin and, thus, protected from cleavage. Thus, plotting the size distribution of sequenced fragments of ATAC Seq libraries would be a good computational approach to address if the low signal to background ratio is due to cell oversaturation with Tn5 transposase.

A recently published Fast-ATAC Seq protocol (Corces *et al.*, 2016), designed for primary blood cells, used 50µl of tagmentation mix for only 5,000 primary haemopoietic cells (including blood $CD4^+$ and $CD8^+$ T cells). With majority of our samples having more than 5,000 cells and use of only 30 µl - 35µl of tagmentation mix made oversaturation less likely. Nevertheless, we believe that for any further analysis, construction of fragment size distribution plots will be crucial.

The loss of reads due to mitochondrial contamination let us question if low sequencing depth has an impact on the high background noise observed. *Ou et al 2018* used already published ATAC Seq data and showed that downsizing the library size to 2.6M uniquely mapped reads did not affect the patterns seen in quality control plots, yet the peak height decreased (Ou *et al.*, 2018). Similar observation was made by *Buenrostro et al 2013* when they compared the assay performance between 50,000 and 500 GM12878 lymphoblastoid cells (Buenrostro *et al.*, 2013). Taken together, both studies showed that low cell counts in combination with low sequencing depth are naturally susceptible to higher background signal.

Another computational experiment to determine the ATAC Seq library quality would be visualization of read distribution around regulatory regions of known housekeeping genes. It is expected for housekeeping genes to be actively transcribed and, thus, their promoter sites should be opened.

In summary, we are not sure what the was reason behind the low signal: noise rations observed. However, after consideration of work done by Corces *et al* 2017 we noticed that $CD4^+$ T cells libraries showed only 10% FRiP, when libraries were prepared by first generation ATAC Seq protocol (Corces *et al.*, 2017). This suggests that the poor FRiP seen is not due to failure of library preparation but an inherit problem with the protocols used.